

**RAQAMLI TEXNOLOGIYALAR VA SUN'IY INTELLEKTNI
RIVOJLANTIRISH ILMIY-TADQIQOT INSTITUTI HUZURIDAGI
ILMIY DARAJALAR BERUVCHI DSc.13/30.12.2021.T.142.01 RAQAMLI
ILMIY KENGASH**

O'ZBEKISTON MILLIY UNIVERSITETI

ALLABERDIYEV BOBUR BAXTIYOROVICH

**O'ZBEK VA QOZOQ TILLARI O'RTASIDA MASHINALI TARJIMA
TIZIMI**

05.01.11 – Raqamli texnologiyalar va sun'iy intellekt

**TEXNIKA FANLARI BO'YICHA FALSAFA DOKTORI (PhD) DISSERTATSIYASI
AVTOREFERATI**

Toshkent – 2024

**Texnika fanlari bo'yicha falsafa doktori (PhD) dissertatsiyasi avtoreferati
mundarijasi**

**Оглавление автореферата диссертации доктора философии (PhD)
по техническим наукам**

**Contents of dissertation abstract of doctor of philosophy (PhD) on technical
sciences**

Allaberdiyev Bobur Baxtiyorovich

O'zbek va qozoq tillari o'rtasida mashinali tarjima tizimi 3

Аллабердиев Бобур Бахтиёрович

Система машинного перевода между узбекским и казахским языками23

Allaberdiev Bobur Bakhtiyorovich

Machine translation system between Uzbek and Kazakh languages 43

E'lon qilingan ishlar ro'uxati

Список опубликованных работ

List of published works 47

**RAQAMLI TEXNOLOGIYALAR VA SUN'IY INTELLEKTNI
RIVOJLANTIRISH ILMIY-TADQIQOT INSTITUTI HUZURIDAGI
ILMIY DARAJALAR BERUVCHI DSc.13/30.12.2021.T.142.01 RAQAMLI
ILMIY KENGASH**

O'ZBEKISTON MILLIY UNIVERSITETI

ALLABERDIYEV BOBUR BAXTIYOROVICH

**O'ZBEK VA QOZOQ TILLARI O'RTASIDA MASHINALI TARJIMA
TIZIMI**

05.01.11 – Raqamli texnologiyalar va sun'iy intellekt

**TEXNIKA FANLARI BO'YICHA FALSAFA DOKTORI (PhD) DISSERTATSIYASI
AVTOREFERATI**

Toshkent – 2024

Texnika fanlari bo'yicha falsafa doktori (PhD) dissertatsiyasi mavzusi O'zbekiston Respublikasi Oliy ta'lim, fan va innovatsiyalar vazirligi huzuridagi Oliy attestatsiya komissiyasida B2023.4.PhD/T4191 raqam bilan ro'yxatga olingan.

Dissertatsiya Mirzo Ulug'bek nomidagi O'zbekiston Milliy universitetida bajarilgan.
Dissertatsiya avtoreferati uch tilda (o'zbek, rus, ingliz (rezyume)) Ilmiy kengashning veb-sahifasida (www.airi.uz) va "Ziyonet" Axborot ta'lim portalida (www.ziyonet.uz) joylashtirilgan.

Ilmiy rahbar:

Raxmonov Zafar Ravshanovich
fizika-matematika fanlari doktori, dotsent

Rasmiy opponentlar:

Muxamediyeva Dilnoz Tulkunovna
texnika fanlari doktori, professor

Madraximov Shavkat Fayzullayevich
texnika fanlari doktori, professor

Yetakchi tashkilot:

Toshkent axborot texnologiyalari universiteti Nukus filiali

Dissertatsiya himoyasi Raqamli texnologiyalar va sun'iy intellektni rivojlantirish ilmiy-tadqiqot instituti huzuridagi DSc.13/30.12.2021.T.142.01 raqamli Ilmiy kengashning 2024-yil 27 ~~dekabr~~ soat 16:00 dagi majlisida bo'lib o'tadi (Manzil: 100125, Toshkent shahri, Mirzo Ulug'bek tumani, Bo'z-2 mavzesi, 17A uy. Tel.: (99871) 263-41-98, e-mail: info@airi.uz).

Dissertatsiya bilan Raqamli texnologiyalar va sun'iy intellektni rivojlantirish ilmiy-tadqiqot instituti Axborot-resurs markazida tanishish mumkin (50 raqam bilan ro'yxatga olingan). (100125, Toshkent shahri, Mirzo Ulug'bek tumani, Bo'z-2 mavzesi, 17A uy. Tel.: (99871) 263-41-98).

Dissertatsiya avtoreferati 2024-yil "17" ~~dekabr~~ kuni tarqatildi.

(2024-yil "10" ~~dekabr~~ dagi 19 raqamli reyestr bayonnomasi).



N.Ravshanov
Ilmiy darajalar beruvchi ilmiy kengash raisi, texnika fanlari doktori, professor

F.M.Nuraliyev
Ilmiy darajalar beruvchi ilmiy kengash ilmiy kotibi, texnika fanlari doktori, professor

Sh.X.Fozilov
Ilmiy darajalar beruvchi ilmiy kengash qoshidagi ilmiy seminar raisi, texnika fanlari doktori, professor

KIRISH (falsafa doktori (PhD) dissertatsiyasining annotatsiyasi)

Dissertatsiya mavzusining dolzarbligi va zarurati. Jahon miqyosida olib borilayotgan ko'plab ilmiy-amaliy tadqiqotlar shuni ko'rsatadiki, internet termog'ida axborotlarni nafaqat tez va samarali izlash, balki turli tillar o'rtasida samarali mashinali tarjima tizimlarini yaratish muhim ahamiyat kasb etmoqda. Axborot texnologiyalarining jadal rivojlanishi va jamiyatni ommaviy raqamlashtirish tufayli insonning kundalik hayotida turli tillardagi axborotlar hajmi kundan-kunga ortib bormoqda. Madaniyatlararo muloqot ko'p odamlar hayotining ajralmas qismidir, shuning uchun tarjima tizimlaridan foydalanish oddiy internet foydalanuvchisi uchun kundalik bajariladigan ishdir. Tillararo murakkablik va noaniqlik barcha tabiiy tillarni qayta ishlash ishlariga xos bo'lgan muammolardan sanaladi. Shu munosabat bilan, til to'siqlarini bartaraf etishga qaratilgan izlanishlarga talab odatdagidan ancha yuqori. Shu sababli sun'iy intellekt texnologiyalari asosida turkiy tillar uchun mashinali tarjima muammosini kompleks tadqiq qilish, turkiy tillar o'rtasida mashinali tarjimini amalga oshirishning samarali modellarini ishlab chiqish va algoritm hamda dasturiy majmuasini ishlab chiqish sun'iy intellekt sohasining muhim vazifalaridan biri bo'lib qolmoqda.

Hozirgi kunda jahonda turkiy tillarni kompyuterda qayta ishlash sohasida ko'plab tadqiqotlar olib borildiki, turkiy tillardan ma'lumot to'plash, qayta ishlash va kibermakonda uzatish tillari sifatida faol foydalanish tizimlari hamda texnologiyalari yaratishga katta e'tibor qaratilmoqda. Bunda tillararo tarjima tizimlariga, matnlarning morfologik tahlili va annotatsiyasini avtomatlashtirish muhim ahamiyat kasb etmoqda. Shu sababli tillarning xususiyatlarini inobatga olgan holda yig'iladigan yirik hajmli ma'lumotlarni tizimlashtirish, ularni qayta ishlash va ular asosida yangi bilimlarni hosil qilishga ehtiyoj ortib bormoqda. Xususan, turkiy tillar uchun neyron mashinali tarjima modelini qurish va parallel korpus yaratish maqsadli ilmiy tadqiqotlardan hisoblanadi.

Mamlakatimizda amaliy va fundamental ahamiyatga ega bo'lgan sanoat, logistika, tibbiyot, qishloq va suv xo'jaligi, telekommunikatsiya va ijtimoiy sohalardagi amaliy masalalarni zamonaviy axborot texnologiyalari asosida hal etishga alohida e'tibor qaratilmoqda. Xususan, matn ko'rinishida yozilgan iqtibos hamda fikrlarni sentiment tahlil qilish, o'zbek tili uchun sentiment tahlil korpusini qurish va uning asosida axborot modellar yaratish, tematik modellashtirish va hujjatlarning mazmunan haqqoniylik parametrini aniqlash bo'yicha salmoqli natijalarga erishildi. "Sun'iy intellektni mashina yordamida o'qitish uchun yirik hajmli davlat tilidagi raqamli ma'lumotlarni shakllantirish, shuningdek, davlat tilidagi nutqni tahlil va sintez qilishni qo'llovchi intellektual dasturiy mahsulotlarni yaratish" ustuvor yo'nalishlar etib belgilangan¹. Qaror ijrosini ta'minlashda turkiy tillar uchun grammatik qoidalarining ontologik modellarini va neyron mashinali tarjima modelini qurish, mashinali o'qitish algoritmilarini hamda dasturiy majmuasini ishlab chiqish muhim ahamiyatga ega.

¹ O'zbekiston Respublikasi Prezidentining 2021-yil 17 fevraldagi "Sun'iy intellekt texnologiyalarini jadal joriy etish uchun shart-sharoitlar yaratish to'g'risida"gi PQ-4996-son qarori.

O'zbekiston Respublikasi Prezidentining 2018-yil 27-apreldagi PQ-3682-son "Innovatsion g'oyalar, texnologiyalar va loyihalarni amaliyotga joriy qilish tizimini yanada takomillashtirish chora-tadbirlari to'g'risidagi" qarori, 2019-yil 24-mayda O'zbekiston Milliy universitetidagi fan va ta'lim vakillari bilan uchrashuvdagi ma'ruzasi, 2019-yil 21-oktyabrdagi "O'zbek tilining davlat tili sifatidagi nufuzi va mavqeyini tubdan oshirish chora-tadbirlari to'g'risida"gi PF-5850-son Farmoni, 2019-yil 8-oktyabrdagi "O'zbekiston Respublikasi oliy ta'lim tizimini 2030-yilgacha rivojlantirish Konsepsiyasini tasdiqlash to'g'risida"gi PF-5847-sonli Farmoni, 2020-yil 5-oktyabrdagi "Raqamli O'zbekiston – 2030" strategiyasini tasdiqlash va uni samarali amalga oshirish chora-tadbirlari to'g'risida"gi PF-6079-son qarori, 2020-yil 20-oktyabrdagi "Mamlakatimizda o'zbek tilini yanada rivojlantirish va til siyosatini takomillashtirish chora-tadbirlari to'g'risida"gi PF-6084-son Farmoni, 2020-yil 29-oktyabrdagi "Ilm-fanni 2030-yilgacha rivojlantirish Konsepsiyasi"ni tasdiqlash to'g'risida"gi PF-6097-son Farmoni, 2020-yil 6-noyabrdagi "O'zbekistonning yangi taraqqiyot davrida ta'lim-tarbiya va ilm-fan sohalarini rivojlantirish chora-tadbirlari to'g'risida"gi PF-6108-son qarori, 2021-yil 17-fevraldagi "Sun'iy intellekt texnologiyalarini jadal joriy etish uchun shart-sharoitlar yaratish to'g'risida"gi PQ-4996-son qarori, 2022-yil 28-yanvardagi "2022-2026-yillarga mo'ljallangan Yangi O'zbekistonning taraqqiyot strategiyasi to'g'risida"gi PF-60-son Farmoni, 2022-yil 6-iyuldagi "2022-2026-yillarda O'zbekiston Respublikasining innovatsion rivojlanish strategiyasini tasdiqlash to'g'risida"gi PF-165-son Farmoni hamda mazkur faoliyatga tegishli boshqa normativ-huquqiy hujjatlarda belgilangan vazifalarni amalga oshirishga ushbu dissertatsiya tadqiqoti muayyan darajada xizmat qiladi.

Tadqiqotning respublika fan va texnologiyalari rivojlanishining ustuvor yo'nalishiga mosligi. Mazkur tadqiqot ishi O'zbekiston Respublikasi fan va texnologiyalar rivojlanishining IV. "Axborotlashtirish va axborot-kommunikatsiya texnologiyalarini rivojlantirish" ustuvor yo'nalishi doirasida bajarilgan.

Muammoning o'rganilganlik darajasi. Jahonda ikki til o'rtasida tarjima tizimi, mashinali tarjima, tarjima modellari bo'yicha bir qator tadqiqotlar olib borgan olimlarning ishlari o'rganilib chiqildi. Jumladan, mashinalardan foydalanish orqali tarjima qilish haqidagi birinchi g'oya Gotfrid Vilgelm Leybnits va Rene Dekartlar tomonidan ilgari surilgan. Keyinchalik AQShda Jorjtaun universitetida Leon E.Doster boshchiligidagi mashinali tarjima bo'yicha birinchi loyiha paydo bo'ldi. Rossiyada O.S.Kulagina va I.A.Melchuk mualliflik qilgan fransuz tilidan rus tiliga mashinali tarjimaning birinchi eksperimental tizimi tomonidan yaratilgan. Qozog'istonda T.K.Kayupov, E.Churikov, A.A.Sharpbaylar qozoq tilidan boshqa tillarga mashinali tarjima tizimlari va algoritmlarini ishlab chiqqan.

Til bo'yicha tarjimon dastur yaratishdan oldin tilning morfologik tahlili, tilda gap tuzilish strukturalari bo'yicha model va algoritmlar yaratilishi talab etiladi. O'zbek tili uchun morfologik tahlil ishlari uzoq vaqt davom etgan va bu bo'yicha juda samarali ishlar olib borilgan hamda olib borilmoqda. Jumladan, A.K.Po'latov

va M.Hakimovlar tomonidan kompyuter tilshunosligining muammolarini ishlab chiqish, mashinali tarjimani yaratishda yuzaga keladigan muammolar, o'zbek tili uchun formal grammatik tasvirlashning matematik modellari, ingliz va o'zbek tillarini solishtirish bo'yicha ilmiy tadqiqotlar olib borilgan. Sh.A.Nazirov tomonidan turkiy tillar uchun elektron lug'atlarni shakllantirish masalasi davlat amaliyoti loyihalari doirasida ishlab chiqarilgan.

M.Aripov, N.Ignatov, Sh.Madraximov, A.Norov, N.Abdurahmonova, U.Tuliyev, G'.Matlatipov va S.Matlatipovlar tomonidan o'zbek tili fonetikasi hamda morfologiyasi sohasida bir qator ilmiy ishlar olib borilmoqda. Ushbu tadqiqotlarda o'zbek tilidagi polisillabik so'zlar uchun avtomat bo'g'in ajratish va ko'rish usullari hamda algoritmlari, lingvo-matematik modellari yaratilgan, o'zbek tili morfologiyasining morfotaktika va morfofonika ilovalari asosida Prolog mantiqiy dasturlash tili yordamida o'rganilgan hamda UzMORPP morfologik instrumenti ishlab chiqilgan. Ushbu yo'nalish bo'yicha o'tkazilgan tadqiqotlar shuni ko'rsatadiki, morfologik tahlil, gap tuzilish strukturalari bo'yicha salmoqli ishlar olib borilgan.

Dissertatsiya tadqiqotining dissertatsiya bajarilgan muassasaning ilmiy tadqiqot ishlari rejalari bilan bog'liqligi. Dissertatsiya tadqiqoti Mirzo Ulug'bek nomidagi O'zbekiston Milliy universitetining ilmiy tadqiqot ishlari rejasiga muvofiq REP-25112021/113 – "UzUDT: "O'zbek tilida tabiiy tilni qayta ishlash uchun universal bog'liqlik daraxti korpusi va uning semantik tahlili" (2021-2024) mavzusidagi ilmiy tadqiqot loyihasi doirasida bajarilgan.

Tadqiqotning maqsadi mashinali tarjimaning an'anaviy usullari va ilg'or modellari asosida 2 ta turkiy tillar o'rtasida tarjima qilish uchun mos algoritmlar hamda dasturiy majmuasini ishlab chiqishdan iborat.

Tadqiqotning vazifalari:

o'zbek va qozoq tillarining grammatikasini o'rganish hamda ular o'rtasidagi o'xshashlik va farqlarni aniqlash;

chog'ishtirayotgan tillar grammatik qoidalaridan foydalanilgan holda qiyosiy ontologik modellarni qurish;

o'zbek-qozoq tilining o'zbek-qozoq tilining qiyosiy ontologik modeli asosida bilimlar bazasi shakllantirish;

o'zbek tili xususiyatlarini inobatga olgan holda neyron mashinali o'qitish algoritmidan foydalanib tarjima modelini yaratish;

neyron to'ri asosida o'zbek va qozoq tillari uchun mashinali tarjima tizimini ishlab chiqish;

tarjima matnning sifatini baholash (BLEU) va o'zaro entropiya metrikalaridan foydalanib, tarjima aniqligi va semantik yaqinligini aniqlash.

Tadqiqot obyekti o'zbek-qozoq tillari uchun mashinali tarjima tizimidan iborat.

Tadqiqot predmeti o'zbek-qozoq parallel korpusi asosida mashinali tarjima tizimini yaratish uchun statistik modellar, neyron mashinali tarjima usullari va algoritmlaridan iborat.

Tadqiqot usullari. Tadqiqot ishida sun'iy intellekt usullari, berilganlarning intellektual tahlili, tasniflash, rasmiylashtirish, ontologik modellashtirish, modulli

va obyektga yo'naltirilgan dasturlash texnologiyalaridan foydalanilgan.

Tadqiqotning ilmiy yangiliklari quyidagilardan iborat:

o'zbek va qozoq tillari grammatik qoidalaridan foydalanilgan holda qiyosiy ontologik modellari qurilgan;

chog'ishtirilayotgan tilning o'zbek-qozoq tilining qiyosiy ontologik modeli asosida bilimlar bazasi shakllantirildi;

o'zbek tili xususiyatlarini inobatga olgan holda neyron mashinali o'qitish algoritmidan foydalanib tarjima modeli yaratilgan;

neyron to'ri asosida o'zbek va qozoq tillari uchun mashinali tarjima tizimi ishlab chiqilgan.

Tadqiqotning amaliy natijalari quyidagilardan iborat:

o'zbek va qozoq tillarining grammatikalari o'rtasidagi o'xshashlik hamda farqlar aniqlangan;

o'zbek va qozoq tillari grammatik qoidalarining ontologik modellari qurilgan; jumlar darajasida o'zbek va qozoq tillari uchun matnlar o'rtasidagi yozishmalar muvofiqlashtirilgan;

o'zbek va qozoq tillari uchun neyron mashinali tarjima tizimi ishlab chiqilgan.

Tadqiqot natijalarining ishonchliligi. Ilmiy tadqiqot natijalarining to'g'riligi va ishonchliligi tadqiqot ishlari natijasida yaratilgan aniq modellar hamda usullarning to'g'riligi, ishlab chiqilgan algoritm va dasturiy majmua orqali olingan natijalarning qiyosiy tahlillari orqali asoslangan.

Tadqiqot natijalarining ilmiy va amaliy ahamiyati.

Tadqiqotining natijalarining ilmiy ahamiyati tabiiy tilning grammatik qoidalarini rasmiylashtirish va kompyuterda qayta ishlash, shuningdek, dasturlashning modulli, obyektga yo'naltirilgan metodologiyalari, tabiiy tilning grammatik qoidalarini rasmiylashtirish, sun'iy intellekt va kompyuter lingvistikasi sohalarida qo'llaniladigan nazariyalarni o'z ichiga oladi.

Tajriba natijasida tavsiya etilgan modellarning yutuqlari solishtirilib, tizimning kuchli va zaif tomonlari aniqlandi. Ishlab chiqilgan modellar va usullar tabiiy tildagi matnlarni avtomatik ravishda abstraksiyalash, qayta ishlashning boshqa amaliy vazifalarini hal qilishda, shuningdek, axborot qidiruv tizimlarida, o'rganish uchun dasturiy ta'minotni ishlab chiqishda, shuningdek, elektron ta'lim tizimlarida, baholashda qo'llanilishi mumkin. Bu esa turkiy tilli davlatlar o'rtasida internet, elektron pochta, ijtimoiy tarmoqlar orqali ijtimoiy va iqtisodiy aloqalarni rivojlantirishga olib keladi. Tadqiqot ishida olib borilgan tajribalar ta'lim, fan, san'at, sog'liqni saqlash va boshqa ijtimoiy sohalaridagi har qanday ma'lumotlarni tarjima qilishda qo'llanilishi bilan izohlanadi.

Tadqiqot natijalarining joriy qilinishi. O'zbek va qozoq tillari o'rtasida mashinali tarjima tizimi uchun ishlab chiqilgan algoritmlar hamda ularning dasturiy ta'minoti yordamida olingan ilmiy yangiliklar asosida:

o'zbek va qozoq tillarining grammatikalari o'rtasidagi aniqlangan o'xshashlik hamda farqlar, ikki til uchun yaratilgan parallel korpus, shakllantirilgan grammatik qoidalarining ontologik modellari va ishlab chiqilgan mashinali o'qitish algoritmlardan Muhammad al-Xorazmiy nomidagi Toshkent axborot

texnologiyalari universiteti Samarqand filialida bajarilgan grantda foydalanilgan (O'zbekiston Respublikasi Raqamli texnologiyalar vazirligining 2024-yil 20-noyabrdagi 33-8/7790-son ma'lumotnomasi). Natijada o'zbek tili milliy korpusining matnlar bazasini hajmini ko'paytirish va dasturiy ta'minotning aniqligini 15 % oshirish imkonini bergan;

o'zbek-qozoq tilining qiyosiy ontologik modeli asosida shakllantirilgan bilimlar bazasi "Unicon SOFT" MCHJ Samarqand viloyat filialida tadbiiq etilgan (O'zbekiston Respublikasi Raqamli texnologiyalar vazirligining 2024-yil 20-noyabrdagi 33-8/7790-son ma'lumotnomasi). Mazkur bilimlar bazasi asosida tuzilgan dasturiy vositaning tarjimon modeli natijalari Google va Yandex tarjimon tizimlari kabi natija berishi hamda ma'lumotlar to'plamini ko'paytirish orqali natijalarni yaxshilash imkoniyatiga erishish mumkinligi aniqlangan. Ma'lumotlarni tarjima qilish uchun sarf qilinadigan vaqt sarfini 11 %ga qisqartirishga xizmat qilgan;

"O'zbektelekom" AK Samarqand shahar telekommunikatsiya bog'lamasida mijozlar bilan aloqa jarayonida sinovdan o'tkazilgan va amaliyotga joriy etilgan. Sinov natijasida tavsiya etilgan modellarning yutuqlari solishtirilib, tizimning kuchli va zaif tomonlari aniqlangan. Ishlab chiqilgan modellar va usullar tabiiy tildagi matnlarni avtomatik ravishda abstraksiyalash hamda qayta ishlashning boshqa amaliy masalalarini hal qilishda, shuningdek, axborot-qidiruv tizimlarida, o'quv dasturlarini ishlab chiqishda va elektron ta'lim tizimlarida foydalanish mumkinligi aniqlandi. Tadqiqot ishida olib borilgan tajribalar ta'lim, fan, san'at, sog'liqni saqlash va boshqa ijtimoiy sohalaridagi har qanday ma'lumotlarni tarjima qilishda qo'llanilishi bilan izohlanadi (O'zbekiston Respublikasi Raqamli texnologiyalar vazirligining 2024-yil 20-noyabrdagi 33-8/7790-son ma'lumotnomasi). Natijada mijozlar bilan muloqot jarayonida 13.5 % tezlik bo'yicha samarali natijaga erishish imkonini taqdim qildi.

Tadqiqot natijalarining aprobatyasi. Dissertatsiyaning asosiy nazariy va amaliy natijalari 6 ta xalqaro hamda 5 ta respublika ilmiy-amaliy anjumanlarida muhokamadan o'tkazilgan.

Tadqiqot natijalarining e'lon qilinganligi. Dissertatsiya mavzusi bo'yicha jami 14 ta ilmiy ish chop etilgan, shulardan O'zbekiston Respublikasi Oliy attestatsiya komissiyasining doktorlik dissertatsiyalari asosiy ilmiy natijalarini chop etish tavsiya etilgan ilmiy nashrlarda 5 ta maqola, jumladan, 3 tasi xorijiy ilmiy jurnallarda nashr etilgan. O'zbekiston Respublikasi Intellektual mulk agentligining EHM uchun dasturiy mahsulot qayd etilganligi haqida 3 ta guvohnoma olingan.

Dissertatsiyaning tuzilishi va hajmi. Dissertatsiya tarkibi kirish, to'rt bob, xulosa, foydalanilgan adabiyotlar ro'yxati va ilovalardan iborat. Dissertatsiyaning hajmi 118 betni tashkil etgan.

DISSERTATSIYANING ASOSIY MAZMUNI

Dissertatsiyaning **Kirish** qismida dissertatsiya mavzusining dolzarbligi va zaruriyati asoslangan, tadqiqotning O'zbekiston Respublikasi fan va texnologiyalar taraqqiyotining ustuvor yo'nalishlariga mosligi ko'rsatilgan. Tadqiqotning maqsad

va vazifalar belgilab olingan hamda tadqiqot obyekti va predmeti aniqlangan, tadqiqotning ilmiy yangiligi va amaliy natijalari bayon qilingan, olingan natijalarning ishonchiligi asoslangan va nazariy hamda amaliy ahamiyati ochib berilgan, tadqiqot natijalari amaliyotga joriy etilganligi, chop etilgan ishlar va dissertatsiya ishining tuzilishi haqida ma'lumotlar keltirib o'tilgan.

Dissertatsiyaning "**Mashinali tarjima masalalari**" deb nomlangan birinchi bobida mashinali tarjima nazariy asoslari, umumiy ko'rinishlari, tarjima qilishda qo'llaniladigan usul va modellar qiyosiy tahlil etilgan. Mashinali tarjima masalalari dunyoning rivojlangan mamlakatlarida qay yo'sinda yechilgan, qanday algoritmlar tavsiya etilgan va qanday natijalarga erishilganligi yoritilgan.

Tarjima jarayonida qo'llaniladigan asosiy usullar keltirilgan:

Qoida asoslangan mashinali tarjima usuli. Qoidalarga asoslangan mashinali tarjima (Rule-based Machine Translation), yoki lingvistik qoidalarga asoslangan mashinali tarjima 1950-yillarda qo'llanilgan birinchi usul edi. Birinchi RBMT tizimlari 1970-yillarning boshlarida ishlab chiqilgan. RBMT asosida Systran, Yaponiya MT tizimi, EUROTRA tizimi Apertium, GramTrans, Promt kabi ilg'or tizimlar paydo bo'ldi.

Bunday tarjima texnologiyasining keng ko'lamli vazifasi tabiiy tillarni ma'lum bir dasturiy ta'minot va apparat kompyuter muhitida amalga oshirilgan obyektlar sifatida rasmiy modellashtirish, shuningdek, turli xil tahliliy protseduralarni modellashtirish, tabiiy tillar uchun til birliklari va matnlarini integratsiya qilish, ya'ni bunday vazifa quyidagi tartiblarni talab qiladi:

- 1) grafik (simvol) tahlil;
- 2) morfologik hamda leksik tahlil va sintez;
- 3) sintaktik tahlil va sintez;
- 4) semantik transformatsiyalar.

Til tuzilishining barcha darajalari MT davomida qayta ishlanadi.

Modelga asoslangan mashinali tarjima. Misol asosidagi mashinali tarjima (Example Based Machine Translation) usuli 1984-yilda yapon tadqiqotchisi M.Nagao tomonidan taklif qilingan. Asosiy g'oya shundan iboratki, agar matnning muayyan turi va tor sohalari uchun ilgari tarjima qilingan iboralarning katta zaxirasi mavjud bo'lsa, keyingi matnlarning tarjimasi inson tarjimasiga o'xshash bo'ladi. Modelga asoslangan mashinali tarjima usulida "misol" yordamida tarjima qilish prinsipi, ikki tilda qo'sh gaplar qurish qo'llaniladi. Foydalanish qoidalariga ko'ra, u "o'xshashlik bo'yicha tarjima" (translation by analogy) deb ataladi.

Tarjima usulining asosiy tahminlari:

– Odamlar chuqur grammatik qoidalarga asoslangan sodda gaplarni tarjima qilmaydi.

– Aksincha, odamlar tarjimada dastlabki qadamlarni asl gapni qandaydir kichik guruhlariga bo'lish orqali qo'yadi. Keyin u ushbu iboralarni maqsadli tilga tarjima qiladi va oxirgi bosqichda ularni birlashtirib, samarali natijalar blokini yaratadi.

– iboralarni tarjima qilishda allaqachon "xotirada saqlangan" ma'lumotlardan foydalangan holda misollar keltiriladi.

Statistik tarjima usulida (SMT) o'nlab yillar davomida tarjima paradigmasi

sifatida hukmronlik qilgan. SMTning amaliy qo'llanilishi, odatda, turli uzunlikdagi so'z va iboralarni tarjima qiladigan frazeologik tizimlar hisoblanadi.

Statistik usullarni o'rganish zarurati Shennonning shovqinli kanal modeliga muvofiq tarjima jarayonini tarjima qilishdir.

Bu fikrga ko'ra, ko'rsatuvchi tildagi H gapi eshittirish kanali orqali o'tgan shovqin ta'sirida o'zgarib, asl tildagi K gapga o'tadi. Usulning prinsipi – translyatsiya kanalining chiqish nuqtasida hisobga olinadigan asl tildagi K jumlasidan kelib chiqib, maqsadli tildagi asl matnga kirish. Bunga erishish uchun quyidagi tenglamaning yechimini topish kifoya:

$$\hat{H} = \underset{H \in \text{Maqsadli til}}{\operatorname{argmax}} P(H|K) \quad (1)$$

Agar (1) tenglamadagi ehtimollik qiymatini Bays qoidasiga muvofiq qayta yozsangiz:

$$\hat{H} = \underset{H \in \text{Maqsadli til}}{\operatorname{argmax}} \frac{P(K|H)P(H)}{P(K)} \quad (2)$$

Bu tenglamada $P(K)$ ehtimollik barcha H ifodalari uchun doimiy bo'lgani uchun argmax operatori uchun natija o'zgarmaydi. Bu holda (2) tenglamani quyidagicha yozish mumkin:

$$\hat{H} = \underset{H \in \text{Maqsadli til}}{\operatorname{argmax}} \underbrace{P(K|H)}_{\text{Tarjima modeli}} \underbrace{P(H)}_{\text{Til modeli}} \quad (3)$$

Ushbu tenglamaning ikkita alohida komponenti mavjud. Birinchisi $P(K|H)$ tarjima modeli deb ataladi va maqsad tildagi H jumlasini K -gapning asl tildagi tarjimasi bo'lishi mumkinligini ko'rsatadi. Ikkinchi komponent $P(H)$ jumlasining maqsad tildagi ehtimolini bildiradi. Bu ikki ehtimollik H gapning qiymatini ko'paytirish natijasidir.

Neyron mashinali tarjima usuli. Neyron Mashinali Tarjima (Neural Machine Translation (NMT)) – integratsiyalashgan modelda butun jumalarni modellashtirish orqali so'zlarni ketma-ketlik ehtimolini bashorat qilish uchun sun'iy neyron tarmog'idan foydalanadigan mashinali tarjima usuli. Chuqur o'rganish birinchi marta 1990-yillarda nutqni aniqlashda qo'llanilgan. 2014-yilda birinchi marta Google-ning iboralarga asoslangan tarjima tizimiga sezilarli yaxshilanishlar taklif qilindi.

NMT an'anaviy mashinali tarjima tizimlarining ko'pgina kamchiliklarini bartaraf etish imkoniyatiga ega bo'lgan istiqbolli yondashuv sifatida taklif qilingan. NMT ning kuchi uning kirish matniga qarab yakuniy o'zgarishlarni bevosita manba matniga o'rganish qobiliyatidadir.

Neural Machine Translationning afzalligi shundaki, u an'anaviy iboraga asoslangan mashinali tarjimaning ko'plab aniq dizayn variantlarini chetlab o'tadi. Biroq amalda NMT tizimlari tarjima tizimlariga qaraganda kamroq aniqroqdir, ayniqsa, mavjud bo'lgan eng yaxshi tarjima tizimlari uchun ishlatiladigan juda katta ma'lumotlar to'plamini o'rganishda. Neyron mashinali tarjimaning uchta kamchiliklari bu kamchiliklar bilan bog'liq: sekin o'qish va sekin chiqish tezligi, unikal so'zlar bilan ishlashning samarasizligi va ba'zan asl gapdagi barcha so'zlarni tarjima qila olmaslik. Birinchidan, NMT tizimini katta tarjima ma'lumotlar bazasida o'qitish, odatda, ko'p vaqt va hisoblash resurslarini talab

qiladi.

Gibrid mashinali tarjima usuli. Gibrid mashinali tarjima – bu mashinali tarjimaning bir tizimida mashinali tarjimaning bir nechta usullarini qo‘llash bilan tavsiflangan mashinali tarjima usuli. Ushbu texnologiya tarjima xotirasi asosida tezda sozlash, uslubni umumlashtirish va juda foydali ma’lumotlarni – terminologik lug‘atlarni olish imkonini beradi.

Dissertatsiyaning “**O‘zbek va qozoq tillarining qiyosiy ontologik modellari**” deb nomlangan ikkinchi bobida ikki til o‘rtasida mashinali tarjima tizimini yaratishda tillarning grammatikasini o‘rganish va ular o‘rtasidagi o‘xshashlik hamda farqlarni aniqlash va grammatik qoidalarining ontologik modellarini qurish amalga oshirilgan.

Ikki til bir-biridan farqli jihatlari ko‘p bo‘lib, hattoki harflarda ham farqlar borligi aniqlandi. 1-jadvalda tillar o‘rtasida unli tovushlar keltirilgan.

1-jadval

O‘zbek va qozoq tillaridagi unli tovushlar

Til	Unli tovushlar soni	Unli tovushlar
O‘zbek	6	a, e, i, o, o‘, u
Qozoq	9	a, ə, e, ы, i, o, ө, ұ, ү

O‘zbek va qozoq tillaridagi har bir so‘z turkumining o‘zgarish tizimini ta’minlovchi o‘ziga xos so‘z yasovchi qo‘shimchalar hamda birikmalarga ega.

Turkiy tillarda eng ko‘p qo‘llaniladigan leksik-semantik sohalardan biri otdir. Otlarga ko‘plik qo‘shimchalari qo‘shilishi ularning birlik va ko‘plik guruhlariga bo‘linishiga asos bo‘ladi. O‘zbek va qozoq tillarining ko‘plik shakllari 2-jadvalda keltirilgan.

2-jadval

O‘zbek va qozoq tillarining ko‘plik shakllari

O‘zbek	Qozoq	O‘zbekcha misol	Qozoqcha misol
-lar, -lari	-лар, -лер	bola-lar	бала-лар, үй-лер
-miz, -imiz		vazifa-miz, kitob-imiz	
-ngiz, -ingiz		vazifa-ngiz, kitob-ingiz	
	-дар, -дер		ғалым-дар, өзен-дер
	-тар, -тер		ат-тар, шөп-тер

Tadqiqot ishimizda qozoq tili o‘zbek tilidan qo‘shimcha va so‘z boyligi bilan farq qilishi aniqlandi. Egalik, so‘z yasovchi qo‘shimchalar va boshqa sohalarda ham ko‘p ma’lumotli til ekanligini natijalarda ko‘rish mumkin.

3-jadvalda o‘zbek va qozoq tillarining egalik qo‘shimchalari keltirilgan.

3-jadval

O‘zbek va qozoq tillarining egalik qo‘shimchalari

Shaxs	O‘zbek	Qozoq	O‘zbek	Qozoq	O‘zbek	Qozoq
	Olmosh		Qo‘shimchalar		Masalan	
1-shaxs birlik	Mening	Менің	-m, -im	-м, -ім, -ым	Mening bolam	Менің балам

2-shaxs birlik	Sening	Сениң	-ng, -ing	-н, -иң, -ың	Sening bolang	Сениң балаң
3-shaxs birlik	Uning	Оның	-i, -si	-ci, -сы, -i, -ы	Uning bolasi	Оның баласы
1-shaxs ko'prik	Bizning	Биздің	-miz, -imiz	-imiz, -ымыз	Bizning bolamiz	Биздің баламыз
2-shaxs ko'prik	Sizning (hurmat)	Сиздің	-ing, -ingiz	-ңыз, -ңиз, -ыңыз, -иңиз	Sizning bolangiz	Сиздің балаңыз
3-shaxs ko'prik	Ularning	Олардың	-i, -si, -lari	-ci, -сы, -i, -ы	Ularning bolalari	Олардың балалары
2-shaxs ko'prik		Сендердің		-н, -иң, -ың		Сендердің балаларың
2-shaxs ko'prik		Сиздердің		-ңыз, -ңиз, -ыңыз, -иңиз		Сиздердің балаларыңыз

Protege platformasi ontologik modellashtirishning ikkita asosiy usulini qo'llab-quvvatlaydi: Protege-Frames muharriridan foydalanish; Protege-OWL muharriridan foydalanish. Protege-da qurilgan ontologiyalar funksional kengaytma modullarining mavjudligi tufayli osongina kengaytiriladigan arxitekturaga ega ochiq ontologiyalar bo'lib, ular RDF (RDFSchem), OWL va XMLSchema kabi ko'plab shakllarga o'tkazilishi mumkin.

1. O'zbek tilidagi sodda gaplarning tarkibiy daraxtini tuzish. Ma'lumki, Xomskiyning formal kontekstsiz grammatikasi (Context-Free Grammar (CFG)) ma'lum til sintaksisini rasman tavsiflash imkonini beruvchi matematik vositadir. Hozirgi vaqtda CFG tabiiy til komponentlari tuzilishini modellashtirish uchun keng qo'llaniladigan rasmiy tizimdir.

Ikki turkiy til sodda gaplarining sintaktik qoidalarini qurish uchun avvaliga Xomskiyning kontekstsiz grammatikasidan foydalanib, ular uchun tarkibiy daraxt tuzdik, so'ngra ularning semantikasini hisobga olgan holda Protege muhitida ontologik model yaratdik.

Ushbu qurilma yordamida ikki o'zbek tildagi sodda gaplarning sintaktik qoidalarini tuzamiz va ularning tarkibiy daraxtini yaratamiz. Lekin buning uchun biz maxsus lingvistik belgilar – teglarni joriy qilishimiz kerak.

O'zbekcha gaplarning tuzilishini tavsiflovchi teglar 4-jadvalda keltirilgan.

4-jadval

O'zbekcha gaplarning tuzilishini tavsiflovchi teglar

Teg	Ingliz	Rus	O'zbek	Qozoq
S	Simple sentence	Простое предложение	Oddiy jumla	Жай сөйлем
Sub	Subject	Подлежащее	Ega	Бастауыш
Obj	Object	Дополнение	To'ldiruvchi	Толықтауыш
Abr	Abreviatura	Определение	Aniqlovchi	Анықтауыш
Adl	Adverbial	Обстоятельство	Hol	Пысықтауыш
Pre	Predicate	Сказуемое	Kesim	Баяндауыш

CFG har biri til belgilarini birlashtirish va joylashtirish usullarini belgilaydigan qoidalar to'plamidan yoki qaror qabul qilish qoidalaridan iborat.

Umumiy G tipidagi CFG quyidagi parametrlar bilan aniqlanadi:

$$G = \langle N_s, T_s, R, S \rangle \quad (4)$$

Bu yerga:

N_s – terminal bo'lmagan belgilar (o'zgaruvchilar) to'plami;

T_s – terminal belgilar to'plami (konstantalar): bu yerda $N_s \cap T_s = \emptyset$;

$R - A \rightarrow \alpha$ turini yechish qoidalari to'plami, bu yerda A - terminal bo'lmagan belgi, $\alpha - N_s \cup T_s$ alifbosidagi belgilar qatori, ya'ni $\alpha \in (N_s \cup T_s)^*$;

S – boshlang'ich terminal bo'lmagan belgi.

Gapning tuzilishi ikki qismda ko'rsatilishi mumkin: ot va fe'l. O'zbekcha sodda jumlar sintaksisini maxsus CFGlar yordamida tasvirlash mumkin.

Masalan, o'zbekchada quyidagi sodda gapni aytaylik:

1. "Samad kitob o'qiyabdi" – "Самад китап оқып отыр";
2. "Samad kutubxonada kitob o'qiyabdi" – "Самад китапханада китап оқып отыр";
3. "Samad onasi bilan kecha keldi" – "Самад анасымен кеше келді";
4. "Eng yaqin do'stim Samad o'qiyabdi" – "Ең жақын досым Самад оқып отыр".

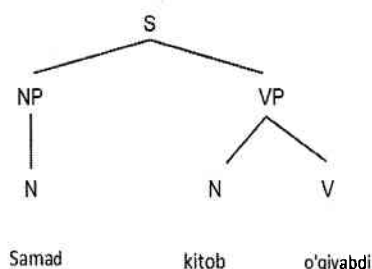
Ushbu jumalarning tuzilishini tavsiflash uchun biz KE parametrlari uchun quyidagi qiymatlarni beramiz:

$$N_s = \{S, NP, VP, Adj, Adv\}$$

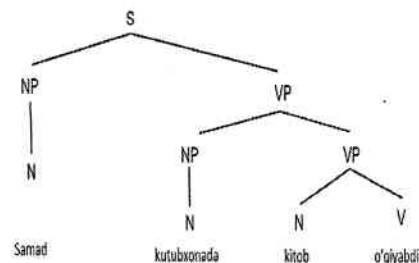
$$T_s = \{S, a, d, e, g, h, i, k, l, m, n, o, p, r, t, s, \$, u, \ddot{u}, y\}$$

$$R = \{S \rightarrow NP \mid VP, NP \rightarrow N \mid N \mid Adj \mid Adv, VP \rightarrow N \mid V \mid Adv \mid NP \mid VP\}$$

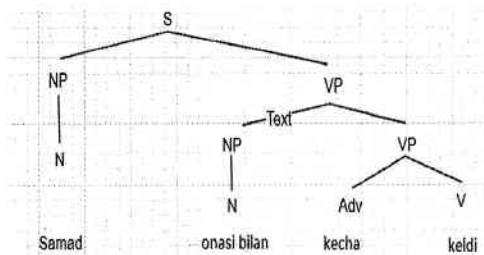
Ushbu grammatik qoidadan foydalanib, yuqorida ko'rsatilgan o'zbek tilidagi jumlar tarkibiy qismlarining tarkibiy daraxtini 1-4-rasmlarda ko'rsatish mumkin:



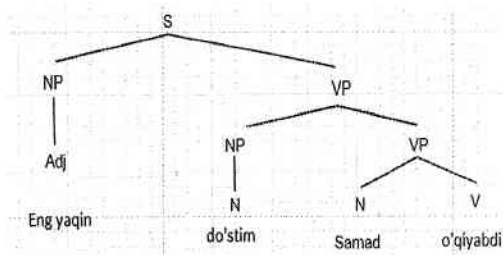
1-rasm. $S(NP(N),VP(N,V))$ ning komponentlar daraxti.



2-rasm. $S(NP(N),NP(N,N)),VP(N,V)$ komponent daraxti.

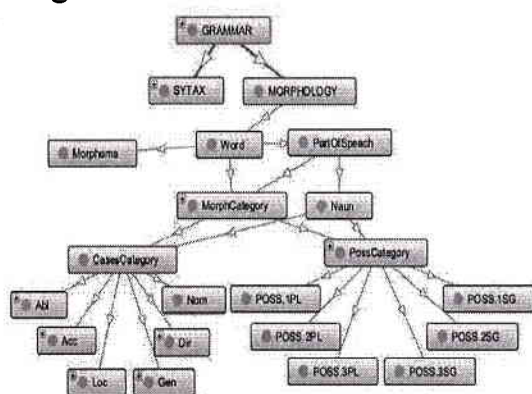


3-rasm. Komponentlar daraxti
S(NP(N), VP(NP(N), VP(Adv,V)).

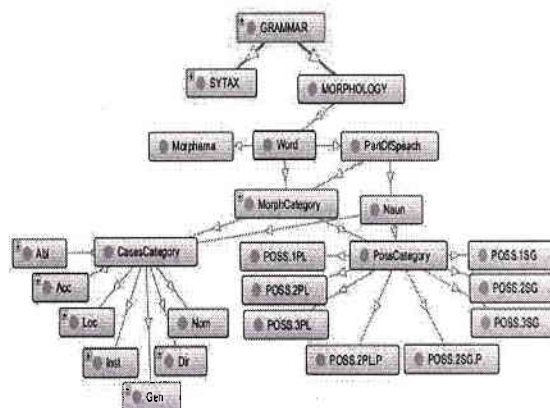


4-rasm. Komponentlar daraxti
S(NP(Adj), VP(NP(N), VP(N,V)).

5-rasmda o'zbek tilidagi otning Protege muhitida yaratilgan ontologik modelining bir qismi, 6-rasmda qozoq tilidagi ot ontologik modelining bir qismi keltirilgan.



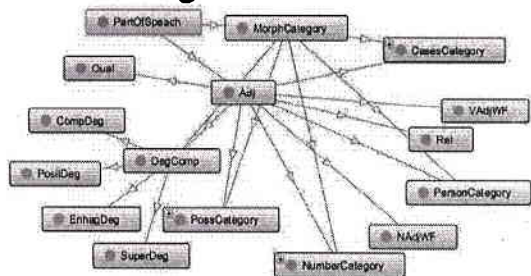
5-rasm. O'zbek tilida otning ontologik modeli.



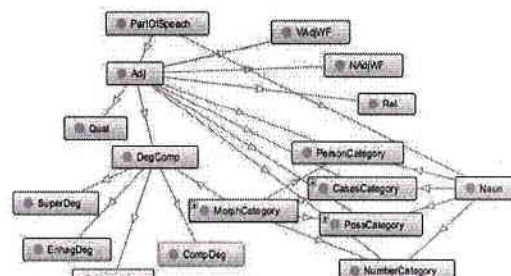
6-rasm. Qozoq tilida otning ontologik modeli.

O'zbek va qozoq tillarida sifatdoshlarning yasalishi bosh sifatlarga kuchaytiruvchi bo'g'in qo'shish orqali yasaladi. O'zbek va qozoq tilida kuchli harf "-" belgisi bilan ajratilib yoziladi. Masalan, o'zbek tilida yap-yangi, sap-sariq, top-toza, qozoq tilida **ап-анық**, **әп-әдемі**, **жеп-жеңіл** va boshqalar.

7-rasmda o'zbek tilidagi sifatning Protege muhitida yaratilgan ontologik modelining bir qismi, qozoq tilidagi sifat ontologik modelining bir qismi esa 8-rasmda keltirilgan.



7-rasm. O'zbek tilida otning ontologik modeli.



8-rasm. Qozoq tilida otning ontologik modeli.

Dissertatsiyaning "O'zbek va qozoq tillari o'rtasida mashinali tarjima tizimi" nomli uchinchi bobida o'zbek-qozoq tillari mashinali tarjima uchun ma'lumotlar to'plamini shakllantirish jarayonlari va neyron mashinali tarjimani

ishlab chiqish yoritilgan.

Parallel korpuslarni yaratish uchun ikki tilda mavjud bo'lgan ma'lumotlar to'plash u yoqda tursin, neyron tarmoq modeli uchun o'quv ma'lumotlari sifatida foydalanish uchun ochiq va yetarlicha katta bo'lgan tayyor matn korpusini to'plash juda qiyin hamda ko'p vaqt talab qiladigan vazifadir, ayniqsa, o'zbek va qozoq tillari kabi kam resursli tillar uchun. Ma'lumotlar to'plamini quyidagi manbalar orqali shakllantirdik:

- Inson huquqlari umumjahon deklaratsiyasi(UDHR);
- Apertium qozoq-o'zbek test ma'lumotlar to'plami;
- COVID19 Myth busters korpusi;
- Veb-yangilik ma'lumotlari (Kazinform sayti)
- Adabiy kitoblar;
- Qo'lda tarjima.

Yaratilgan parallel jumalarning statistik ma'lumotlarining batafsil tavsifi 5-jadvalda keltirilgan.

5-jadval

Parallel jumalarning statistik ma'lumotlari

№	Bosqich	Manba nomi	Gaplar soni	So'zlar (uz)	So'zlar (kz)	Unikal so'zlar soni (uz)	Unikal so'zlar soni (kz)	O'rtacha jumladagi so'zlar soni (uz)	O'rtacha jumladagi so'zlar soni (kz)
1.	1	UDHR	61	1380	1535	724	760	22.6	25.2
2.		Apertium qozoq-o'zbek test ma'lumotlari	49	321	345	206	219	6.6	7.0
3.		COVID19 Myth busters korpusi	28	288	314	197	209	10.3	11.2
4.	2	Veb-yangilik ma'lumotlari	315	4220	4335	2331	2348	13.4	13.8
5.		Kichkina shahzoda	1688	10870	12270	4200	4633	6.4	7.3
6.		Kapitan qizi	3008	26767	28276	9753	11010	8.9	9.4
7.		O'n besh yoshda kapitan	8012	81094	89637	18549	20284	10.1	11.2
8.		Shayton ko'prigi	11822	99319	110259	24655	29454	8.4	9.3
9.	3	Qo'lda tarjimalar	100000	526207	536674	47355	48768	5.3	5.4
Umumiy soni / o'rtacha			124983	750466	783645	107970	117685	10.2	11.1

Yakuniy bosqich barcha manbalarni yagona ma'lumotlar to'plami sifatida birlashtirish va takroriy jumalarni olib tashlash uchun oxirgi filtrni qo'llashdan iborat bo'lgan olingan kichik to'plamlarga qo'llanildi. Yakuniy birlashtirilgan ma'lumotlar to'plamining statistikasi 6-jadvalda keltirilgan.

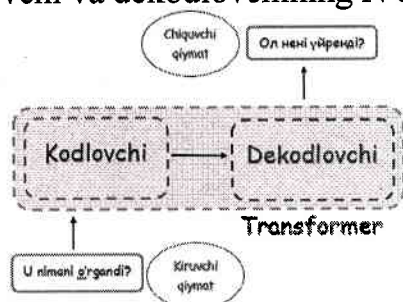
6-jadval

Yakuniy birlashtirilgan ma'lumotlar to'plamining statistikasi

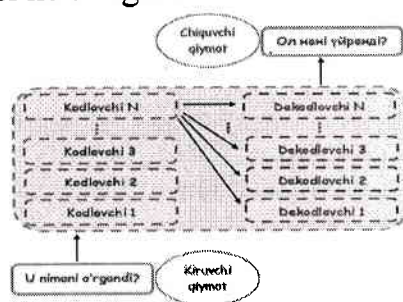
Har bir til uchun statistika	O'zbek	Qozoq
Gaplarning umumiy soni	121138	
So'zlarning umumiy soni	746738	759357
Unikal so'zlarning umumiy soni	89967	104888
Har bir jumladagi so'zlarning o'rtacha soni	6.2	6.3

Ikki til uchun ma'lumotlar to'plamidan foydalanib tarjimon model qurishni boshlaymiz. Kam resursli tillar uchun eng mos tarjimon usullaridan biri bu neyron mashinali tarjima (NMT) hisoblanadi. Hozirgi kunda ommalashgan NMT modellaridan biri bu transformer modeli hisoblanadi.

Transformerlar zamonaviy NLP modeli bo'lib, kodlovchi-dekodlovchi arxitekturasining evolyutsiyasi hisoblanadi. Elementlar o'rtasidagi munosabatlarni tahlil qilish orqali kontekst va ma'noni tushunish uchun maxsus ishlab chiqilgan va buning uchun deyarli butunlay matematik texnikaga tayanadi. 9-rasmda Transformer modelining kodlovchi va dekodlovchi strukturasi keltirilgan. Bu qismlar N ta qatlamlardan tashkil topgan struktura orqali ishlaydi. 10-rasmda kodlovchi va dekodlovchining N qatlamli strukturasi keltirilgan.

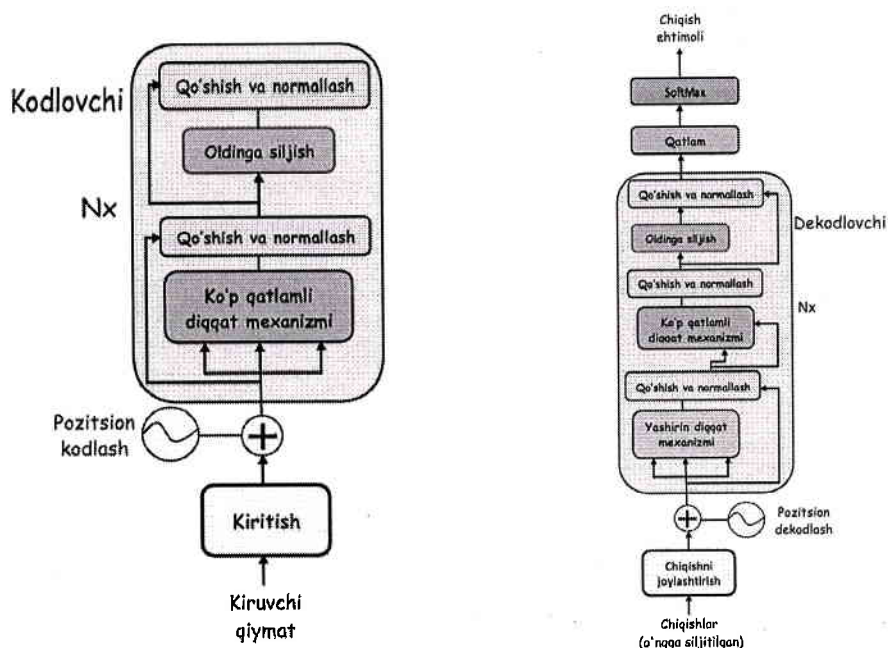


9-rasm. Transformer modeli.



10-rasm. N qatlamli struktura.

Kodlovchining asosiy vazifasi kirish tokenlarini kontekstli tasvirlarga aylantirishdir. Tokenlarni mustaqil ravishda qayta ishlaydigan oldingi modellardan farqli o'laroq, Transformer kodlovchisi har bir tokenning kontekstini butun ketma-ketlikka nisbatan qamrab oladi. Dekodlovchining roli matn ketma-ketligini yaratishga qaratilgan. Kodlovchini aks ettirgan holda, dekodlovchi shunga o'xshash pastki qatlamlar to'plami bilan jihozlangan. U ikkita ko'p yo'nalishli e'tibor qatlamiga, oldinga yo'naltirilgan qatlamga ega va har bir pastki qatlamdan keyin qoldiq ulanishlarni hamda qatlamni normallashtirishni o'z ichiga oladi. 11-rasmda kodlovchi arxitekturasini 12-rasmda dekodlovchi arxitekturasini keltirilgan.

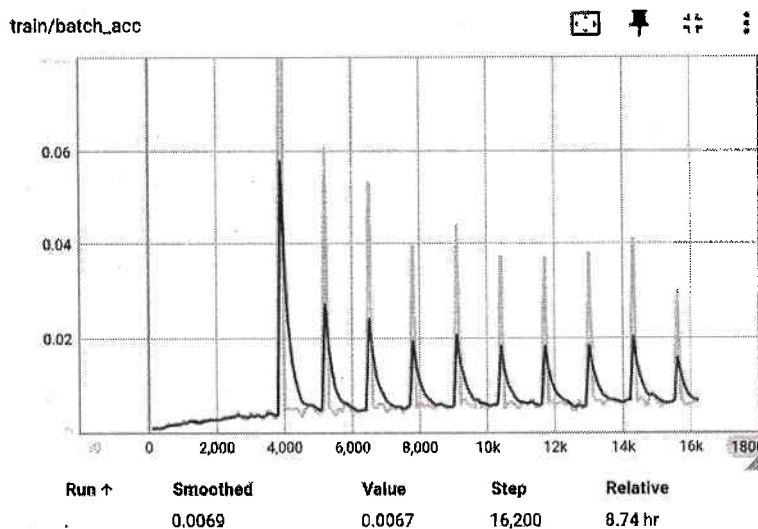


11-rasm. Kodlovchi (Encoder). 12-rasm. Dekodlovchi (Decoder).

Dekodlovchi bitta qatlam bilan chegaralanib qolmaydi. U N ta qatlam bilan tuzilishi mumkin, ularning har biri kodlovchi va undan oldingi qatlamlardan olingan kirishga asoslanadi. Bunday ko'p qatlamli yondashuv modelning bashorat qilish qobiliyatini sezilarli darajada oshirishi mumkin, chunki u turli xil diqqat kombinatsiyalarini yanada nozikroq tushunishni imkonini beradi.

Dissertatsiyaning **“Tarjimon modelining ishonchliligi va natijaviyligi”** nomli to'rtinchi bobida ishonchlilik ko'rsatkichlari, tarjimon modelining natijaviyligi va boshqa tarjimon tizimlari bilan solishtirishdan olingan natijalar yoritilgan.

Birinchi navbatda, ikki til o'rtasidagi mutanosiblik qay darajada ekanligi baholanadi. Gap yasovchi qo'shimchalar, ko'plik shakllari, gap tuzilish strukturalari va boshqalar. 13-rasmda ikki til o'rtasidagi mutanosiblik grafigi keltirilgan.



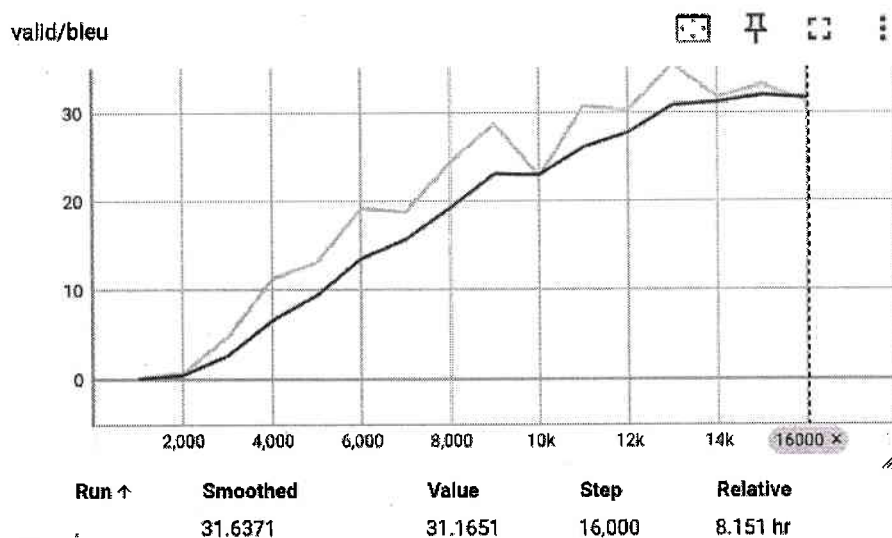
13-rasm. Ikki til o'rtasidagi mutanosiblik grafigi.

BLEU (Bilingual Evaluation Understudy) ballidan foydalanib, NMT chiqishini bir yoki bir nechta mos yozuvlar tarjimalari bilan solishtiradi. Ushbu ko'rsatkich mashinali tarjima va mos yozuvlar matnlari o'rtasidagi n-grammlarning bir-biriga mos kelishini o'lchaydi. BLEU algoritmi inson va mashinali tarjimani solishtirish hamda umumiy parchalarni qidirish asosida tarjimani 0 dan 100 gacha bo'lgan shkalada baholaydi. Asosiy g'oya shundan iboratki, qancha mos kelsa, tarjima shunchalik yaxshi bo'ladi.

BLUE ball quyidagicha hisoblanadi:

$$BLEU = BP \times \exp\left(\sum_{n=1}^N w_n \log p_n\right) \quad (5)$$

BLUE ball hisoblash grafigi 14-rasmda keltirilgan.

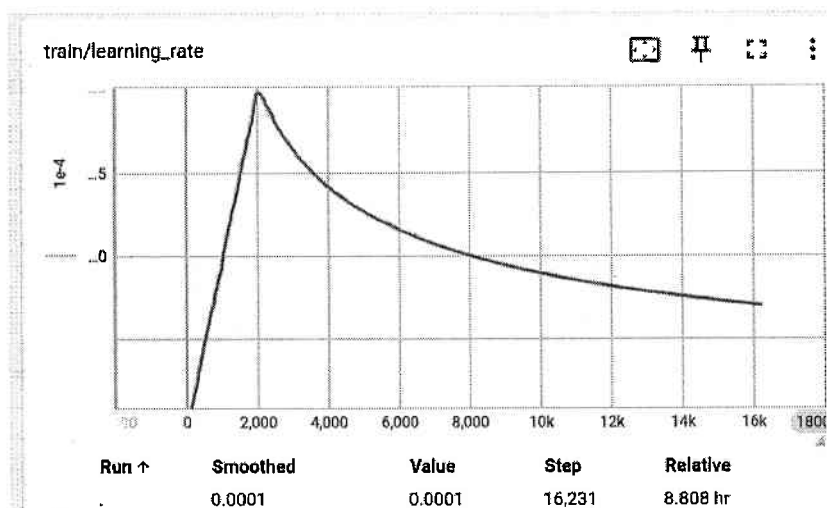


14-rasm. BLUE ball hisoblash grafigi.

Mashinali tarjima sifatini baholash uchun ko'rsatkich. Ko'rsatkich n-gramlardan foydalanishga asoslangan va manba matnni izchil hamda to'g'ri baholashga qaratilgan. BLEU ko'rsatkichidan farqli o'laroq, bu ko'rsatkich aniq so'zlarni moslashtirish bilan birga sinonimlarni moslashtirishdan foydalanadi. Ushbu ko'rsatkich mashhurroq BLEU ko'rsatkichida topilgan muammolarni hal qilish va ibora hamda jumlar darajasida ekspert baholari bilan yaxshiroq bog'lanish uchun mo'ljallangan.

TER (Tarjimani tahrirlash tezligi): gipotezani havolalardan biriga o'zgartirish uchun zarur bo'lgan tahrirlar sonini hisoblab chiqadi.

Tarjimani tahrirlash tezligi grafigi 15-rasmda keltirilgan.

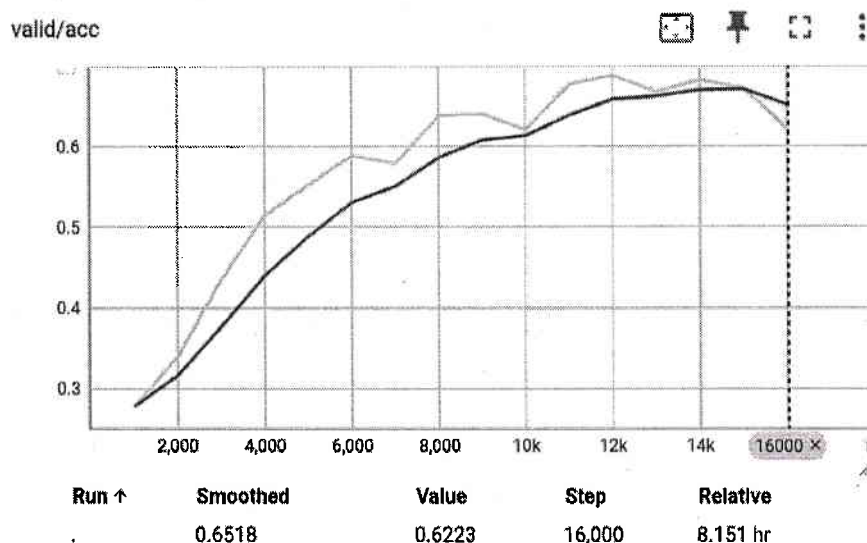


15-rasm. Tarjimani tahrirlash tezligi grafigi.

Ideal holda, tekshirishning aniqligi treningning rivojlanishi bilan ortib borishi kerak, bu model ko‘rinmas ma’lumotlarni yaxshiroq o‘rganish va umumlashtirishni ko‘rsatadi.

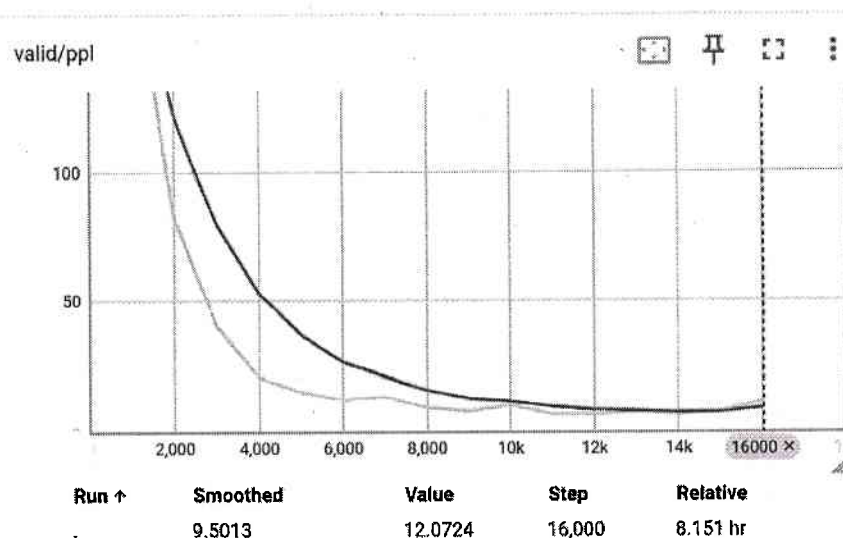
Doimiy o‘sish tendensiyasi samarali o‘rganishni taklif qiladi.

16-rasmda tasdiqlash aniqligi tendensiyasining grafigi keltirilgan.



16-rasm. Tasdiqlash aniqligi tendensiyasi grafigi.

Tabiiy tilni qayta ishlash vazifalarida, ayniqsa, tilni modellashtirish va mashinali tarjimada, chalkashlik (PPL) modelning ishlashini baholash uchun ishlatiladigan umumiy ko‘rsatkichdir. Murakkablik – ehtimollik taqsimoti yoki ehtimollik modeli namunani qanchalik yaxshi bashorat qilishining o‘lchovidir. 17-rasmda chalkashliklarni bartaraf etish grafigi keltirilgan.



17-rasm. Chalkashliklarni bartaraf etish grafigi keltirilgan.

Tarjimon modelini testlash jarayonida har bir qadamda parallel korpus tarkibidagi ma'lumotlarga mosligini tekshirib boradi. Har qadamda tarjima sifati yaxshilanib boradi.

18-rasmda o'zbek-qozoq NMT modelini testlash keltirilgan.



18-rasm. O'zbek-qozoq NMT modelini testlash.

Tarjimon modeli ikki til uchun qilingan bo'lib o'zbek tilidan qozoq tiliga va qozoq tilidan o'zbek tiliga tarjima qilib beradi. Bu ish birinchi marotaba qilingani uchun tarjima modelimizni boshqa ish bilan solishtirish imkoni bo'lmadi. Shu sababli hozirgi kundagi ommalashgan tarjimon tizimlari bilan solishtirdik.

7-jadvalda o'zbek-qozoq va qozoq-o'zbek tarjimon modeldan olingan natijalar keltirilgan.

7-jadval

Tarjimon modeldan olingan natijalar

O'zbek tilidagi gap	Yandex tarjimon	Google tarjimon	NMT modeli (UzKz)
Unga bera oladigan yordamingiz uchun oldindan rahmat aytmogchiman.	Мен оған барлық көмегіңіз үшін алдын ала алғыс айтқым келеді.	Оған кез келген көмегіңіз үшін сізге алдын ала алғыс айтқым	Мен оған көрсете алатын кез келген көмегіңіз үшін сізге алдын ала алғыс

		келеді.	айтқым келеді.
O'qituvchi kirib kelganida gapirishni to'xtating.	Мұғалім кірген кезде сөйлесуді тоқтатыңыз.	Мұғалім кірген кезде сөйлесуді тоқтатыңыз.	Мұғалім кірген кезде сөйлесуді тоқтатыңыз.
Qozoq tilidagi gap	Yandex tarjimon	Google tarjimon	NMT modeli (KzUz)
Ата-анасы сәбиге Акиёши деген ат қойды.	Ota-onalar chaqaloq Akiyoshi nomidagi.	Ota-onalar chaqaloqqa Akiyoshi deb nom berishdi.	Ota-onalar chaqaloqqa Akiyoshi deb nom berishdi.
2008 yilda Delta va Northwest aviakompaniyalari birlashib, dunyodagi eng yirik aviakompaniyani tashkil qilishdi.	2008 жылы Delta және Northwest бірігіп, әлемдегі ең ірі авиакомпанияны құрды.	2008 жылы Delta және Northwest Airlines бірігіп, әлемдегі ең ірі әуе компаниясын құрады.	2008 жылы Delta және Northwest әуе компаниялары бірігіп, әлемдегі ең ірі әуе компаниясын құрады.

Natijalardan ko'rinadiki, tarjimon modelimiz natijalari Google va Yandex tarjimon tizimlari kabi natija bermoqda. Ma'lumotlar to'plamini ko'paytirish orqali undanda yaxshi natijalarga erishish mumkin. Buni bizning tarjimon modelimiz ta'minlaydi.

XULOSA

“O'zbek va qozoq tillari o'rtasida mashinali tarjima tizimi” mavzusidagi dissertatsiya ishi doirasida quyidagi asosiy natijalar olindi.

1. O'zbek va qozoq tillarining grammatikasini o'rganildi hamda ular o'rtasidagi o'xshashlik va farqlarni aniqlandi. Oddiy jumlar sintaksisini rasmiylashtirish uchun Xomskiyning kontekstsiz grammatikasidan foydalanilgan va komponentlar daraxti tuzilgan.

2. O'zbek va qozoq tillari grammatik qoidalaridan foydalanilgan holda qiyosiy ontologik modellari qurildi. Jumlar darajasida o'zbek va qozoq tillari uchun matnlar o'rtasidagi yozishmalar muvofiqlashtirildi.

3. O'zbek-qozoq tilining parallel korpusidan foydalanilgan holda lingvistik qoidalar asosida bilimlar bazasini shakllantirildi. Yaratilgan bilimlar bazasini miqdori 125 ming jumlani tashkil qiladi. Bilimlar bazasi “Hugging Face” elektron platformasiga joylashtirildi.

4. O'zbek-qozoq tillari uchun neyron mashinali tarjimasi ishlab chiqildi. Mashinali tarjima ishlab chiqishda “Transformer” modelidan foydalanildi. Bu model o'zbek tilidan qozoq tiliga va qozoq tilidan o'zbek tiliga tarjima qiluvchi tarjima tizimini yaratish imkonini beradi.

5. Neyron to'ri asosida o'zbek va qozoq tillari uchun mashinali tarjima tizimini ishlab chiqildi. Tarjima tizimi soha ekspertlari uchun qaror qabul qilishda yordamchi sifatida ishlatilib, yakuniy qaror ishonchligini baholash imkoniyatini yuzaga keltirdi.

6. Tarjima matnning sifatini baholash (BLEU) va o'zaro entropiya metrikalaridan foydalanib, tarjima aniqligi va semantik yaqinligi aniqlandi.

**НАУЧНЫЙ СОВЕТ DSc.13/30.12.2021.Т.142.01
ПО ПРИСУЖДЕНИЮ УЧЁНЫХ СТЕПЕНЕЙ ПРИ НАУЧНО-
ИССЛЕДОВАТЕЛЬСКОМ ИНСТИТУТЕ РАЗВИТИЯ ЦИФРОВЫХ
ТЕХНОЛОГИЙ И ИСКУССТВЕННОГО ИНТЕЛЛЕКТА**

НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ УЗБЕКИСТАНА

АЛЛАБЕРДИЕВ БОБУР БАХТИЯРОВИЧ

**СИСТЕМА МАШИННОГО ПЕРЕВОДА МЕЖДУ УЗБЕКСКИМ И
КАЗАХСКИМ ЯЗЫКАМИ**

05.01.11 – Цифровые технологии и искусственный интеллект

**АВТОРЕФЕРАТ ДИССЕРТАЦИИ
ДОКТОРА ФИЛОСОФИИ (PhD) ПО ТЕХНИЧЕСКИМ НАУКАМ**

Ташкент – 2024

Тема диссертации на соискание степени доктора философии (PhD) по техническим наукам зарегистрирована под номером B2023.4.PhD/T4191 в Высшей аттестационной комиссии при Министерстве высшего образования, науки и инноваций Республики Узбекистан.

Диссертация выполнена в Национальном Университете Узбекистана имени Мирзо Улугбека. Автореферат диссертации на трех языках (узбекском, русском, английском (резюме)) размещен на веб-странице Научного совета (www.airi.uz) и на информационно-образовательном портале "Ziyonet" (www.ziyonet.uz).

Научный руководитель:

Рахмонов Зафар Равшанович
Доктор физико-математических наук, доцент

Официальные оппоненты:

Мухамедиева Дилноз Тулкуновна
доктор технических наук, профессор

Мадрахимов Шавкат Файзуллаевич
доктор технических наук, профессор

Ведущая организация:

**Нукусский филиал Ташкентского
университета информационных технологий**

Защита диссертации состоится 11.12.2024 года в 16:00 часов на заседании Научного совета DSc.13/30.12.2021.T.142.01 при Научно-исследовательском институте развития цифровых технологий и искусственного интеллекта. (Адрес: 100125, город Ташкент, Мирзо-Улугбекский район, Буз-2 массив, дом 17А. Тел.: (99871) 263-41-98, e-mail: info@airi.uz)

С диссертацией можно ознакомиться в Информационно-ресурсном центре Научно-исследовательского института развития цифровых технологий и искусственного интеллекта. (регистрационный номер № 50). (Адрес: 100125, г. Ташкент, Мирзо-Улугбекский р-н, Буз-2, 17А. Тел.: (99871) 263-41-98.

Автореферат диссертации разослан «17» декабрь 2024 года.
(реестр протокола рассылки № 19 от «10» декабрь 2024 г.).



Н. Равшанов

Председатель научного совета,
доктор технических наук, профессор.

Ф.М. Нуралиев

Ученый секретарь научного совета,
доктор технических наук, профессор.

Ш.Х. Фазилов

Руководитель научного семинара
при научном совете, доктор
технических наук, профессор.

ВВЕДЕНИЕ (аннотация докторской диссертации PhD)

Актуальность и необходимость темы диссертации. В мировом масштабе множество научно-практических исследований показывают, что не только быстрый и эффективный поиск информации в сети интернет является важным, но и создание эффективных машинных систем перевода между различными языками. Благодаря быстрому развитию информационных технологий и общественной цифровизации объем информации на различных языках в повседневной жизни человека постоянно увеличивается. Межкультурное взаимодействие является неотъемлемой частью жизни многих людей, поэтому использование систем перевода становится обыденной задачей для рядового пользователя интернета.

Сложность и неоднозначность перевода между всеми естественными языками рассматривается как одна из основных проблем, связанных с переработкой языков. В связи с этим спрос на исследования, направленные на нейтрализацию языковых аномалий, обычно довольно высок. Именно поэтому исследование проблемы машинного перевода для тюркских языков на основе технологий искусственного интеллекта, разработка эффективных моделей и алгоритмов, а также создание программных комплексов являются одной из важных задач в области искусственного интеллекта.

В настоящее время в мире проводится множество исследований в области компьютерной обработки тюркских языков, большое внимание уделяется созданию систем и технологий активного использования тюркских языков как языков сбора, обработки и передачи информации. в киберпространстве. В этом случае важное значение приобретают системы межъязыкового перевода, автоматизация морфологического анализа и аннотирования текстов. Поэтому растет потребность систематизировать крупномасштабные данные, собранные с учетом особенностей языков, обработать их и создать на их основе новые знания. В частности, целевыми научными исследованиями являются построение модели нейронного машинного перевода для турецких языков и создание параллельного корпуса.

В нашей стране особое внимание уделяется решению практических и фундаментальных проблем в таких отраслях, как промышленность, логистика, медицина, сельское и водное хозяйство, телекоммуникации и общественные сферы, с применением современных информационных технологий. Одним из важных направлений является «Создание интеллектуальных программных продуктов, способных обрабатывать большие объемы цифровых данных на государственном языке с использованием искусственного интеллекта для обучения». В частности, это включает в себя формирование крупных наборов данных на государственном языке для обучения и создания интеллектуальных программ, которые могут анализировать и синтезировать речь на этом языке. Особенно были достигнуты значимые результаты в анализе текстового материала и выявлении сентиментов, создании корпуса для сентимент-анализа на узбекском языке, разработке информационных моделей на его основе,

тематическом моделировании и определении достоверности содержания документов². Значимыми работами для обеспечения выполнения указа, являются создание онтологических моделей для грамматических правил турецких языков и моделей машинного перевода на их основе, разработка алгоритмов машинного обучения и программных комплексов.

Диссертационное исследование в значительной степени вносит свой вклад в выполнение указанных задач, определенных в указах и других нормативно-правовых актах, связанных с развитием образования и науки в Узбекистане, таких как Постановление Президента Республики Узбекистан от 27 апреля 2018 года № УП-3682 «О мерах по дальнейшему совершенствованию системы внедрения инновационных инициатив, технологий и проектов», выступление на совещании с представителями науки и образования в Национальном университете Узбекистана 24 мая 2019 года, 21 октября 2019 года, Постановление № УП-5850 «О мерах по коренному повышению престижа и статуса узбекского языка как государственного», Постановление № УП-5847 от 8 октября 2019 года «Об утверждении Концепции развития системы высшего образования Республики Узбекистан до 2030 года», Постановление от 24 мая 2019 года № ПП-5850 «О важности улучшения статуса и влияния узбекского языка как государственного языка», Постановление № УП-6079 от 5 октября 2020 года «Об утверждении Стратегии «Цифровой Узбекистан – 2030» и мер по ее эффективной реализации» от 20 октября 2020 года, Постановление № УП-6084 «О мерах по дальнейшее развитие и совершенствование языковой политики в нашей стране, Постановление № УП-6097 от 29 октября 2020 года «Об утверждении концепции развития науки до 2030 года», Постановление № УП-6108 от 6 ноября 2020 года «О мерах по развитию сферы образования и науки в период нового развития Узбекистана», Постановление № УП-4996 от 17 февраля 2021 года «О создании условий для ускоренного внедрения технологий искусственного интеллекта», указ №ПФ-60 от 28 января 2022 года «О Стратегии развития нового Узбекистана на 2022-2026 годы» Указ № УП-165 от 6 июля 2022 года «О стратегии развития Нового Узбекистана на 2022-2026 годы» и других нормативных правовых документах, связанных с данной деятельностью.

Соответствие исследования приоритетным направлениям развития науки и технологий республики. Данное исследование выполнено в соответствии с приоритетным направлением развития науки и технологий Республики Узбекистан IV. «Развитие информатизации и информационно-коммуникационных технологий».

Степень изученности проблемы. Изучены работы ученых, проводивших ряд исследований по системе перевода между двумя языками, машинному переводу и моделям перевода. В частности, первую идею перевода с машин на человека выдвинули Готфрид Вильгельм Лейбниц и Рене Декарт. В США первый план машинного перевода появился под

² O'zbekiston Respublikasi Prezidentining 2021 yil 17 fevraldagi «Sun'iy intellekt texnologiyalarini jadal joriy etish uchun shart-sharoitlar yaratish to'g'risida»gi PQ-4996-son qarori.

руководством Леона Э. Достера в Джорджтаунском университете. Первая экспериментальная система машинного перевода с французского языка на русский авторства О.С. Кулагиной и И.А. Мельчука принадлежит России. В Казахстане Т.К. Каюпов, Е. Чуриков, А.А. Sharpbay разработала системы и алгоритмы машинного перевода с казахского языка на другие языки.

Перед созданием программ перевода по языку требуется анализ морфологии языка, моделирование структурных особенностей языка и разработка соответствующих алгоритмов. Работы по морфологическому анализу узбекского языка продолжают уже длительное время и продолжают быть очень продуктивными. В частности, А.К.Полатов и М.Хакимов провели научные исследования по разработке проблем компьютерной лингвистики, проблем, возникающих при создании машинного перевода, математических моделей формального грамматического представления для узбекского языка, сравнения английского и узбекского языков. Вопрос создания электронных словарей турецкого языка разрабатывался Ш.А.Назировым в рамках проектов государственной практики.

М.Арипов, Н.Игнатов, Ш.Мадрахимов, А.Норов, Н.Абдурахмонова, У.Тулиев, Г.Матлатипов и С.Матлатиповы провели ряд научных исследований в области фонетики и морфологии узбекского языка. В рамках этих исследований были разработаны методы и алгоритмы автоматического распознавания и разделения полисиллабических слов в узбекском языке, а также созданы лингво-математические модели. Были изучены и реализованы методы морфотактики и морфофоники узбекской морфологии с использованием логического программирования на языке Prolog, а также разработан морфологический инструмент UzMORPP. Указанные исследования по этому направлению свидетельствуют о проведении качественных работ в области морфологического анализа и структурного анализа предложений. Однако модели, алгоритмы и программные комплексы для машинного перевода не были исследованы.

Связь темы диссертации с планами научно-исследовательских работ научно-исследовательского учреждения. Данная диссертация выполнена в рамках перспективного плана научно-исследовательской работы Национального Университета Узбекистана имени Мирзо Улутбека на тему REP-25112021/113-“UzUDT «Универсальный корпус дерева зависимостей для обработки естественного языка на узбекском языке и его семантический анализ» (2021-2024).

Цель исследования заключается в разработке соответствующей алгоритма и программного комплекса для перевода между двумя тюркскими языками на основе традиционных методов и передовых моделей машинного перевода.

Задачи исследования:

- освоение грамматики узбекского и казахского языков, а также выявление их сходств и различий;
- построение сравнительных онтологических моделей с использованием

грамматических правил узбекского и казахского языков;

с использованием онтологических моделей узбекско-казахских языков сформирована база знаний;

создать модель перевода с использованием алгоритма нейронного машинного обучения с учетом особенностей узбекского языка;

разработка системы машинного перевода для узбекского и казахского языков на основе нейронной сети;

определить точность перевода и семантическую близость с помощью оценки качества переведенного текста (BLEU) и показателей взаимной энтропии.

Объектом исследования система машинного перевода для узбекского и казахского языков.

Предмет исследования статистические модели, методы нейронного машинного перевода и алгоритмы на основе узбекско-казахского параллельного корпуса.

Метод исследования. В исследовательской работе использованы методы искусственного интеллекта, интеллектуального анализа данных, классификации, формализации, онтологического моделирования, модульных и объектно-ориентированных технологий программирования.

Научная новизна исследования заключается в следующем:

сравнительные онтологические модели построены с использованием грамматических правил узбекского и казахского языков;

с использованием онтологических моделей узбекско-казахских языков сформирована база знаний;

с учетом особенностей узбекского языка создана модель перевода с использованием алгоритма нейронного машинного обучения;

разработана система машинного перевода для узбекского и казахского языков на основе нейронной сети.

Практические результаты исследования заключаются в следующем:

выявлены сходства и различия между грамматиками узбекского и казахского языков;

построены онтологические модели грамматических правил узбекского и казахского языков;

на уровне предложений согласованы соответствия текстов для узбекского и казахского языков;

разработана система нейронного машинного перевода для узбекского и казахского языков.

Достоверность результатов исследования. Научная обоснованность и надежность результатов исследования подтверждены точностью созданных моделей и методов, а также сравнительным анализом результатов, полученных с помощью разработанных алгоритмов и программного комплекса.

Научная и практическая значимость результатов исследований.

Научная значимость результатов исследования включает в себя формализацию грамматических правил естественного языка и их обработку

на компьютере, а также модульные и объектно-ориентированные методологии программирования, теории, применяемые в области искусственного интеллекта и компьютерной лингвистики.

Достижения предложенных моделей были сравнены в ходе эксперимента, что позволило выявить сильные и слабые стороны системы. Разработанные модели и методы могут быть использованы для автоматической абстракции текстов на естественном языке, решения других практических задач обработки, а также в информационно-поисковых системах, разработке программного обеспечения для обучения, электронных образовательных системах и оценки. Это, в свою очередь, способствует развитию социальных и экономических связей между тюркоязычными государствами через интернет, электронную почту и социальные сети. Эксперименты, проведенные в ходе исследования, могут быть применены для перевода информации в таких областях, как образование, наука, искусство, здравоохранение и другие социальные сферы.

Внедрение результатов исследования. На основе научных новшеств, полученных с помощью алгоритмов и программного обеспечения, разработанных для системы машинного перевода между узбекским и казахским языками:

выявление сходств и различий между грамматикой узбекского и казахского языков, создание параллельного корпуса для этих двух языков, разработка онтологических моделей грамматических правил и алгоритмов машинного обучения были реализованы в рамках проекта Самаркандского филиала Ташкентского университета информационных технологий имени Мухаммада аль-Хорезми (справка Министерства цифровых технологий Республики Узбекистан от 20 ноября 2024 года, № 33-8/7790). В результате удалось увеличить объем текстовой базы национального корпуса узбекского языка и повысить точность программного обеспечения на 15%;

база знаний, созданная на основе сравнительной онтологической модели узбекско-казахского языка, была успешно внедрена в Самаркандском областном филиале ООО «Unicon SOFT» (справка Министерства цифровых технологий Республики Узбекистан от 20 ноября 2024 года № 33-8/7790). Проведенный анализ показал, что результаты перевода программного обеспечения, построенного на основе этой базы знаний, сопоставимы с результатами переводческих систем Google и Яндекс. При этом установлено, что дальнейшее улучшение качества перевода возможно за счет увеличения объема данных. Внедрение данной технологии позволило сократить время, необходимое для перевода данных, на 11%;

АО «Узбектелеком» было протестировано и внедрено в процессе установления связи с абонентами Самаркандской городской сети связи. В результате испытания были сопоставлены достижения предложенных моделей и выявлены сильные и слабые стороны системы. Установлено, что разработанные модели и методы могут быть использованы при решении других практических задач автоматического абстрагирования и обработки текстов на естественном языке, а также в системах поиска информации,

разработке учебных программ и системах электронного образования. Эксперименты, проводимые в научно-исследовательской работе, объясняются применением любой передачи информации в образовании, науке, искусстве, здравоохранении и других социальных сферах (справка Министерства цифровых технологий Республики Узбекистан от 20 ноября 2024 года, № 33-8/7790). В результате это дало возможность добиться эффективного результата в процессе общения с клиентами на 13,5% скорости.

Апробация результатов исследования. Результаты исследований обсуждались на 6 международных и 5 республиканской научно-практической конференции.

Публикация результатов исследования. По теме диссертации опубликовано 14 научных работ, в том числе 5 статьи в научных изданиях ВАК Республики Узбекистан, рекомендованных к публикации основных научных результатов докторских диссертаций, в том числе 3 статьи в зарубежных научных журналах, также получены 3 свидетельства об официальной регистрации программы для ЭВМ.

Структура и объем диссертации. Диссертация состоит из введения, четырех глав, заключения, списка использованной литературы и приложений. Объем диссертации состоит из 118 страниц.

ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во введении обосновываются актуальность и востребованность темы диссертации в соответствии с приоритетными направлениями развития науки и технологий Республики Узбекистан, формулируются цель и задачи, а также объект и предмет исследования, изложены научная новизна и практические результаты исследования, обоснована достоверность полученных результатов, раскрыта теоретическая и практическая значимость полученных результатов, приведены перечень внедрений в практику результатов исследования, сведения об опубликованных работах и структура диссертации.

В первой главе диссертации, озаглавленной «**Проблемы машинного перевода**», представлен сравнительный анализ теоретических основ машинного перевода, обзоры, методы и модели, используемые при переводе. Освещено, как решались проблемы машинного перевода в развитых странах мира, какие алгоритмы были рекомендованы и какие результаты были достигнуты.

Представлены основные методы, используемые в процессе перевода:

Метод машинного перевода, основанный на правилах. Машинный перевод на основе правил (Rule-based Machine Translation), или машинный перевод, основанный на лингвистических правилах, был первым методом, использованным в 1950-х годах. Первые системы RBMT были разработаны в начале 1970-х годов. На базе RBMT появились такие современные системы, как Systran, японская система MT, система EUROTRA, Apertium, Wikiwand и Prompt.

Широкой задачей такой технологии перевода является формальное

моделирование естественных языков как объектов, реализованных в конкретной программно-аппаратной компьютерной среде, а также моделирование различных аналитических процедур, интеграции языковых единиц и текстов для естественных языков, т.е. такая задача требует следующих процедур:

- 1) графический (символьный) анализ;
- 2) морфологический и лексический анализ и синтез;
- 3) синтаксический анализ и синтез;
- 4) смысловые трансформации.

При машинном переводе обрабатываются все уровни языковой структуры.

Машинный перевод на основе моделей. Метод машинного перевода на основе примеров (Example Based Machine Translation) был предложен японским исследователем М. Нагао в 1984 году. Основная идея заключается в том, что при наличии большого запаса ранее переведенных фраз для конкретного типа текста и узких областей перевод последующих текстов будет аналогичен человеческому переводу. В методе машинного перевода на основе моделей используется принцип перевода по «примеру» и построение двойных предложений на двух языках. По правилам использования это называется «перевод по аналогии».

Основные предположения метода перевода:

- Люди не переводят простые предложения, основанные на глубоких грамматических правилах.

- Напротив, первые шаги в переводе люди делают, разбивая исходное предложение на несколько небольших групп. Затем он переводит эти фразы на целевой язык и объединяет их на последнем этапе для создания эффективного блока результатов.

- приведены примеры перевода фраз с использованием информации, уже «заложеной в памяти».

Статистический метод перевода (СМТ) доминировал в качестве парадигмы перевода на протяжении десятилетий. Практическим применением SMT обычно являются фразеологические системы, переводящие слова и фразы разной длины.

Необходимость изучения статистических методов заключается в том, чтобы перевести процесс перевода в соответствии с моделью шумного канала Шеннона.

Согласно этой идее, речь H на опорном языке меняется на речь K на языке оригинала под воздействием шума, проходящего через канал вещания. Принцип метода заключается в доступе к исходному тексту на целевом языке из предложения K на языке оригинала, которое рассматривается на выходе канала вещания. Для этого достаточно найти решение следующего уравнения:

$$\hat{H} = \underset{H \in \text{Целевой язык}}{\operatorname{argmax}} P(H|K) \quad (1)$$

Если переписать значение вероятности в уравнении (1) по правилу

Байеса:

$$\hat{H} = \underset{H \in \text{Целевой язык}}{\operatorname{argmax}} \frac{P(K|H)P(H)}{P(K)} \quad (2)$$

В этом уравнении, поскольку вероятность $P(K)$ постоянна для всех выражений H , результат оператора argmax не меняется. В этом случае уравнение (2) можно записать следующим образом:

$$\hat{H} = \underset{H \in \text{Целевой язык}}{\operatorname{argmax}} \underbrace{P(K|H)}_{\text{Модель перевода}} \underbrace{P(H)}_{\text{Модель языка}} \quad (3)$$

В этом уравнении есть два различных компонента. Первая называется моделью перевода $P(K|H)$ и указывает, что предложение H на целевом языке может быть переводом предложения K на исходном языке. Второй компонент представляет вероятность предложения H на целевом языке. Эти две вероятности являются результатом умножения значения утверждения H .

Метод нейронного машинного перевода. Нейронный машинный перевод (Neural Machine Translation NMT) — это метод машинного перевода, который использует искусственную нейронную сеть для прогнозирования вероятности последовательностей слов путем моделирования целых предложений в интегрированной модели. Глубокое обучение впервые было использовано для распознавания речи в 1990-х годах. Впервые в 2014 году в систему фразового перевода Google были внесены существенные улучшения.

NMT был предложен как многообещающий подход, способный преодолеть многие недостатки традиционных систем машинного перевода. Сила NMT заключается в его способности запоминать окончательные изменения на основе входного текста непосредственно в исходном тексте.

Преимущество нейронного машинного перевода заключается в том, что он обходит многие очевидные конструктивные решения традиционного машинного перевода на основе фраз. Однако на практике системы NMT менее точны, чем системы перевода, особенно при изучении очень больших наборов данных, используемых для лучших доступных систем перевода. С этими недостатками связаны три недостатка нейронного машинного перевода: медленное чтение и медленная скорость вывода, неэффективная обработка редких слов и иногда невозможность перевести все слова в исходном предложении. Во-первых, обучение системы NMT на большой базе данных переводов обычно требует много времени и вычислительных ресурсов.

Гибридный метод машинного перевода. Гибридный машинный перевод — метод машинного перевода, характеризующийся использованием нескольких методов машинного перевода в одной системе машинного перевода. Эта технология позволяет осуществлять быструю настройку на основе памяти переводов, обобщение стиля и получение очень полезной информации — терминологических словарей.

Во второй главе диссертации под названием «**Сравнительные онтологические модели узбекского и казахского языков**» рассматриваются грамматика языков и выявление сходств и различий между двумя языками, а также построение онтологических моделей грамматических

правил. , были проведены при создании системы машинного перевода между двумя языками.

Два языка во многом отличаются друг от друга, и было обнаружено, что различия есть даже в буквах. В таблице 1 перечислены гласные звуки разных языков.

Таблица 1

Гласные в узбекском и казахском языках

Язык	Число гласных букв	Гласные буквы
Узбекский	6	а, е, і, о, о', и
Казахский	9	а, ә, е, ы, і, о, ө, ұ, ү

Каждая группа слов в узбекском и казахском языках имеет свои словообразовательные аффиксы и соединения, обеспечивающие систему изменения.

Имя существительное является одним из наиболее используемых лексико-семантических полей в турецком языке. Присоединение суффиксов множественного числа к существительным является основой их разделения на группы единственного и множественного числа. Формы множественного числа узбекского и казахского языков приведены в таблице 2.

Таблица 2

Формы множественного числа узбекского и казахского языков.

Узбекский	Казахский	Узбекский пример	Казахский пример
-lar, -lari	-лар, -лер	bola-lar	бала-лар, үй-лер
-miz, -imiz		vazifa-miz, kitob-imiz	
-ngiz, -ingiz		vazifa-ngiz, kitob-ingiz	
	-дар, -дер		ғалым-дар, өзен-дер
	-тар, -тер		ат-тар, шөп-тер

В ходе нашего исследования было установлено, что казахский язык отличается от узбекского по суффиксам и словарному запасу. По результатам видно, что это очень информативный язык с точки зрения владения, словообразовательных суффиксов и других областей.

В таблице 3 приведены притяжательные суффиксы узбекского и казахского языков.

Таблица 3

Узбекские и казахские притяжательные суффиксы

Лицо	Узбекский	Казахский	Узбекский	Казахский	Узбекский	Казахский
	Местоимение		Окончания		Пример	
1-е лицо единственного числа	Mening	Менің	-m, -im	-м, -ім, -ым	Mening bolam	Менің балам
2-е лицо единственного числа	Sening	Сенің	-ng, -ing	-н, -ін, -ың	Sening bolang	Сенің балаң

3-е лицо единственного числа	Uning	Оның	-i, -si	-ci, -сы, -i, -ы	Uning bolasi	Оның баласы
1-е лицо множественного числа	Bizning	Біздің	-miz, -imiz	-imiz, -ымыз	Bizning bolamiz	Біздің баламыз
2-е лицо множественного числа	Sizning (уважен ие)	Сіздің	-ing, -ingiz	-ңыз, -ңіз, -ыңыз, -іңіз	Sizning bolangiz	Сіздің балаңыз
3-е лицо множественного числа	Ularning	Олардың	-i, -si, -lari	-ci, -сы, -i, -ы	Ularning bolalari	Олардың балалары
2-е лицо множественного числа		Сендердің		-ң, -ің, -ың		Сендердің балаларың
2-е лицо множественного числа		Сіздердің		-ңыз, -ңіз, -ыңыз, -іңіз		Сіздердің балаларыңыз

Платформа Protege поддерживает два основных метода моделирования онтологий: с помощью редактора Protege-Frames; Использование редактора Protege-OWL. Онтологии, построенные в Protege, представляют собой открытые онтологии с легко расширяемой архитектурой благодаря наличию функциональных модулей расширения, которые можно преобразовать во многие формы, такие как RDF (RDFSchema), OWL и XMLSchema.

Создание структурного дерева простых предложений на узбекском языке. Известно, что формальная бесконтекстная грамматика Хомского (Context-Free Grammar (CFG)) представляет собой математический инструмент, позволяющий формально описать синтаксис данного языка. В настоящее время CFG является широко используемой формальной системой моделирования структуры компонентов естественного языка.

Для построения синтаксических правил простых предложений двух тюркских языков мы сначала использовали контекстно-свободную грамматику Хомского для построения для них структурного дерева, а затем создали онтологическую модель в среде Protege с учетом их семантики.

С помощью этого устройства мы составим синтаксические правила простых предложений двух узбекских языков и создадим их структуру. Но для этого нам необходимо ввести специальные лингвистические символы – теги.

Теги, описывающие структуру узбекских предложений, приведены в таблице 4.

Таблица 4

Теги, описывающие структуру узбекских предложений

Тэг	Английский	Русский	Узбекский	Казахский
S	Simple sentence	Простое предложение	Oddiy jumla	Жай сөйлем
Sub	Subject	Подлежащее	Ega	Бастауыш
Obj	Object	Дополнение	To'ldiruvchi	Толықтауыш
Abr	Abreviatura	Определение	Aniqlovchi	Анықтауыш
Adl	Adverbial	Обстоятельство	Hol	Пысықтауыш
Pre	Predicate	Сказуемое	Kesim	Баяндауыш

CFG состоит из набора правил или правил принятия решений, каждое из которых определяет способ объединения и расположения языковых символов.

Общий CFG G-типа определяется следующими параметрами:

$$G = \langle N_s, T_s, R, S \rangle \quad (4)$$

Здесь:

N_s – набор нетерминальных символов (переменных);

T_s – набор символов терминала (константы): здесь $N_s \cap T_s = \emptyset$;

$R - A \rightarrow \alpha$ набор правил решения типа, где A – нетерминальный символ, α – строка символов алфавита $N_s \cup T_s$, т.е. $\alpha \in (N_s \cup T_s)^*$;

S – начальный нетерминальный символ;

Структура предложения может быть представлена двумя частями: существительным и глаголом. Синтаксис простых предложений узбекского языка можно описать с помощью специальных CFG.

Приведем следующее простое предложение на узбекском языке:

1. "Samad kitob o'qiyabdi" – "Самад китап оқып отыр";
2. "Samad kutubxonada kitob o'qiyabdi" – "Самад китапханада китап оқып отыр";
3. "Samad onasi bilan kecha keldi" – "Самад анасымен кеше келді";
4. "Eng yaqin do'stim Samad o'qiyabdi" – "Ең жақын досым Самад оқып отыр";

Для описания структуры этих предложений приведем следующие значения параметров KE:

$$N_s = \{S, NP, VP, Adj, Adv\} \quad T_s = \{S, a, d, e, g, h, i, k, l, m, n, o, p, r, t, s, \$, u, \ddot{u}, y\}$$

$$R = \{S \rightarrow NP \mid VP, NP \rightarrow N \mid Adj \mid Adv, VP \rightarrow N \mid V \mid Adv \mid NP \mid VP\}$$

Используя это грамматическое правило, структурное дерево компонентов приведенных выше узбекских предложений можно представить на рисунках 1-4:

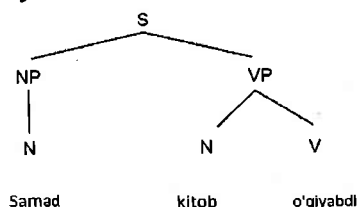


Рисунок 1 – Дерево компонентов $S(NP(N), VP(N, V))$.

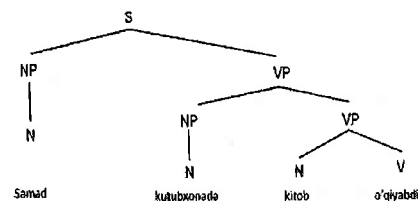


Рисунок 2 – $S(NP(N), NP(N, N)), VP(N, V)$ дерево компонентов

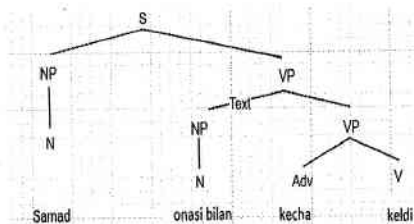


Рисунок 3 – Дерево компонентов S(NP(N), VP(NP(N), VP(Adv,V)))

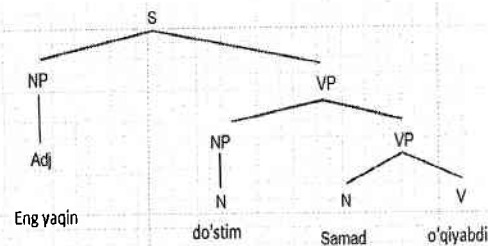


Рисунок 4 – Дерево компонентов S(NP(Adj), VP(NP(N), VP(N,V)))

На рисунке 5 представлена часть онтологической модели существительного в узбекском языке, созданная в среде Protege, на рисунке 6 показана часть онтологической модели существительного в казахском языке.

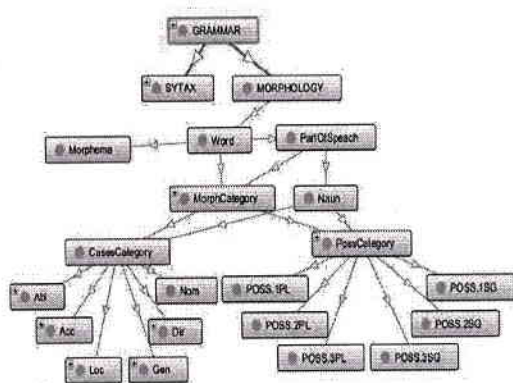


Рисунок 5. Онтологическая модель существительного в узбекском языке



Рисунок 6. Онтологическая модель имени существительного в казахском языке

в узбекском и казахском языках прилагательные образуются путем присоединения к основному прилагательному слога-усилителя. В узбекском и казахском языках пишется сильная буква, разделенная знаком «-». Например, на узбекском языке уар-уангi, сар-сарик, топ-тоза, на казахском ап-анык, эп-эдемі, жеп-жеңіл и другие.

На рисунке 7 представлена часть онтологической модели прилагательного в узбекском языке, созданная в среде Protege, а часть онтологической модели прилагательного в казахском языке представлена на рисунке 8.

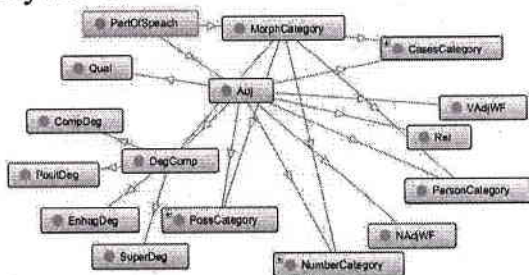


Рисунок 7. Онтологическая модель существительного в узбекском языке

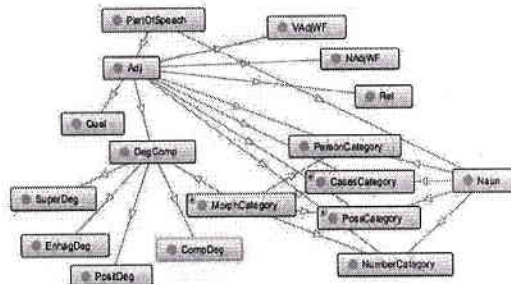


Рисунок 8. Онтологическая модель имени существительного в казахском языке

В третьей главе диссертации под названием «Система машинного перевода между узбекским и казахским языками» описаны процессы формирования набора данных для узбекско-казахского машинного перевода и развитие нейронного машинного перевода.

Сбор готовых текстовых корпусов, которые являются открытыми и достаточно большими, чтобы их можно было использовать в качестве обучающих данных для модели нейронной сети, не говоря уже о сборе данных, доступных на двух языках, для создания параллельных корпусов, является чрезвычайно сложной и трудоемкой задачей. , особенно для языков с низкими ресурсами, таких как узбекский и казахский. Мы создали набор данных из следующих источников:

- Всеобщая декларация прав человека (ВДПЧ);
- Набор данных казахско-узбекского теста Apertium;
- Корпус COVID19 Myth busters;
- Информация в сети новостей (сайт Казинформ)
- Литературные книги;
- Ручной перевод.

Подробное описание статистических данных созданных параллельных предложений представлено в таблице 5.

Таблица 5

Статистика параллельных предложений

№	Этап	Название источника	Слова (узб)	Слова (кз)	Количество уни-Кальных слов (уз)	Количество уни-Кальных слов (кз)	Среднее количество слов в предложении (уз)	Среднее количество слов в предложении (уз)	Среднее количество слов в предложении (kz)
1.	1	UDHR	61	1380	1535	724	760	22.6	25.2
2.		Данные казахстанско-узбекского теста Apertium	49	321	345	206	219	6.6	7.0
3.		Корпус COVID19 Myth busters	28	288	314	197	209	10.3	11.2
4.	2	Данные веб-новостей	315	4220	4335	2331	2348	13.4	13.8
5.		Маленький принц	1688	10870	12270	4200	4633	6.4	7.3
6.		Дочь капитана	3008	26767	28276	9753	11010	8.9	9.4
7.		Капитан в 15 лет	8012	81094	89637	18549	20284	10.1	11.2
8.		Мост дьявола	11822	99319	110259	24655	29454	8.4	9.3
9.	3	Ручные переводы	100000	526207	536674	47355	48768	5.3	5.4
Общее количество/среднее			124983	750466	783645	107970	117685	10.2	11.1

Давайте начнем строить модель перевода, используя набор данных для

двух языков. Одним из наиболее подходящих методов перевода для малоресурсных языков является нейронный машинный перевод (NMT). Одной из популярных сегодня моделей NMT является модель-трансформер.

Трансформеры — это современная модель НЛП, являющаяся развитием архитектуры кодера-декодера. Он разработан специально для понимания контекста и значения путем анализа взаимосвязей между элементами и почти полностью полагается на математические методы. На рисунке 9 показана структура кодирования и декодирования модели Transformer. Эти части работают через структуру, состоящую из N слоев. На рисунке 10 показана N-уровневая структура кодера и декодера.

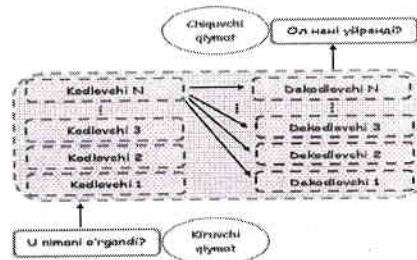
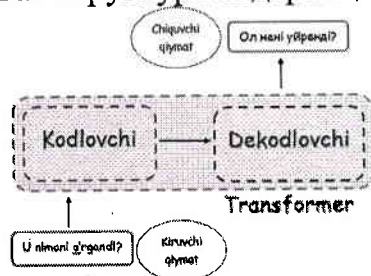


Рисунок 9. Модель трансформера. Рисунок 10. N-слойная структура.

Основная задача кодировщика — преобразовать входные токены в контекстные изображения. В отличие от предыдущих моделей, которые обрабатывают токены независимо, кодировщик Transformer фиксирует контекст каждого токена относительно всей последовательности. Роль декодера заключается в создании текстовых последовательностей. Зеркально отражая кодер, декодер оснащен аналогичным набором подуровней. Он имеет два фокусных слоя многолучевого распространения, прямой уровень, а также включает остаточные соединения и нормализацию уровня после каждого подуровня. На рисунке 11 показана архитектура кодера, а на рисунке 12 показана архитектура декодера.

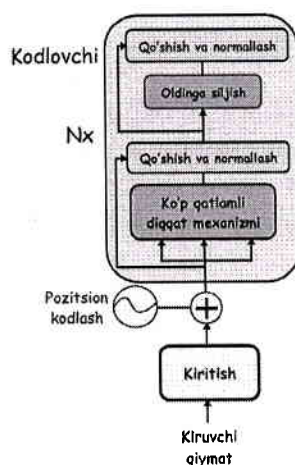


Рисунок 11. Кодер(Encoder)



Рисунок 12. Декодер(Decoder)

Декодер не ограничен одним слоем. Он может состоять из N слоев, каждый из которых основан на входных данных от кодера и предыдущих

слоев. Такой многоуровневый подход может значительно улучшить предсказательную силу модели, поскольку обеспечивает более детальное понимание различных комбинаций внимания.

Четвертая глава диссертации под названием «Надежность и производительность модели транслятора» посвящена показателям надежности, производительности модели транслятора и результатам, полученным при сравнении с другими системами транслятора.

Прежде всего оценивается степень совместимости двух языков. Образующие предложения наречия, формы множественного числа, структуры предложений и т. д. На рисунке 13 показан график корреляции между двумя языками.

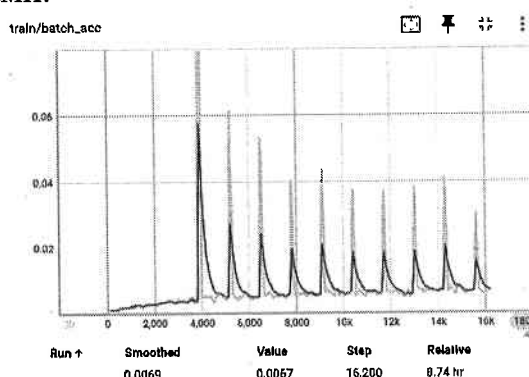


Рисунок 13. График корреляции между двумя языками

Сравнивает выходные данные NMT с одним или несколькими эталонными переводами, используя оценку BLEU (Bilingual Evaluation Understudy). Этот показатель измеряет перекрытие n-грамм между машинным переводом и справочными текстами. Алгоритм BLEU оценивает перевод по шкале от 0 до 100 на основе сравнения человеческого и машинного перевода и поиска общих отрывков. Основная идея заключается в том, что чем больше совпадений, тем лучше перевод.

BLUE балл рассчитывается следующим образом.:

$$BLEU = BP \times \exp\left(\sum_{n=1}^N w_n \log p_n\right) \quad (5)$$

График расчета баллов BLUE представлен на рисунке 14.

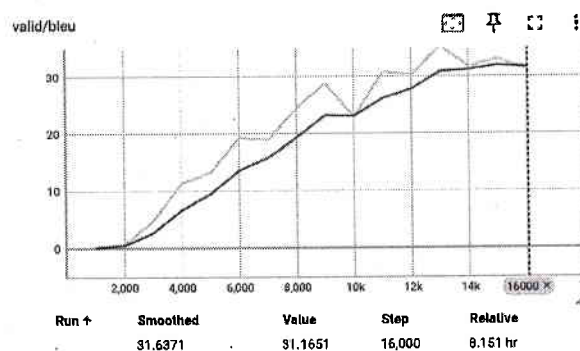


Рисунок 14. График вычисления BLUE балла

Показатель оценки качества машинного перевода. Индикатор основан на использовании n-грамм и направлен на последовательную и правильную оценку исходного текста. В отличие от метрики BLEU, эта метрика использует сопоставление синонимов наряду с точным соответствием слов. Эта метрика предназначена для решения проблем, обнаруженных в более популярной метрике BLEU, и для лучшей связи с экспертными суждениями на уровне фраз и предложений.

TER (скорость редактирования перевода): рассчитывает количество правок, необходимых для изменения гипотезы одной из ссылок.

График скорости редактирования перевода представлен на рисунке 15.

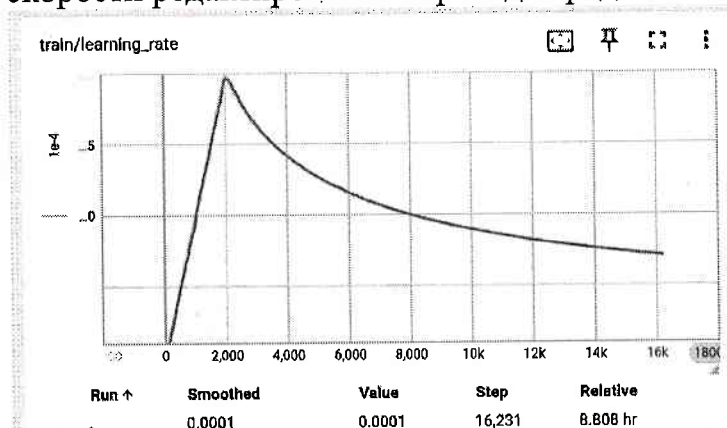


Рисунок 15. График скорости редактирования перевода

В идеале точность проверки должна повышаться по мере прохождения обучения, что указывает на лучшее обучение и обобщение невидимой информации.

Постоянная тенденция к росту предполагает эффективное обучение.

На рис. 16 показан график тенденции точности подтверждения.

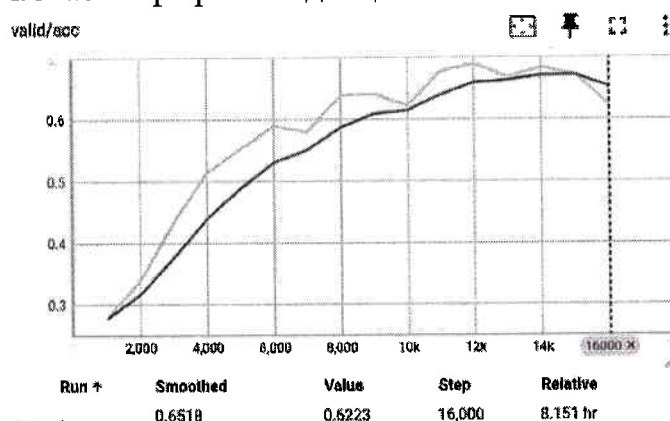


Рисунок 16. График тенденции точности проверки

В задачах обработки естественного языка, особенно языкового моделирования и машинного перевода, путаница (PPL) является распространенным показателем, используемым для оценки производительности модели. Сложность — это мера того, насколько хорошо распределение вероятностей или вероятностная модель предсказывает выборку. На рисунке 17 показан график деконволюции.

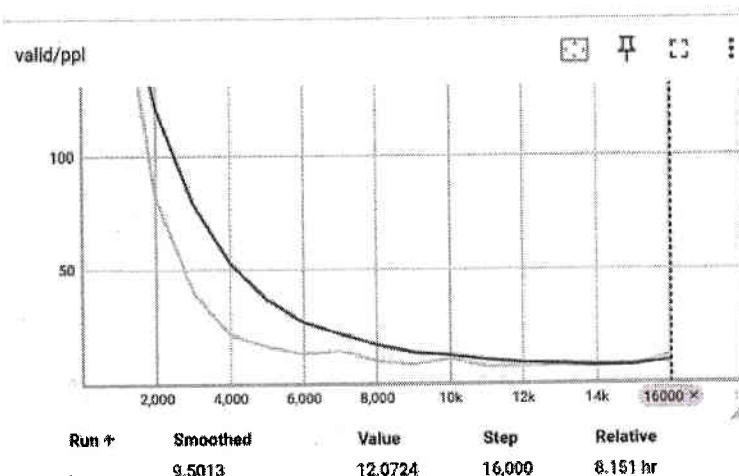


Рисунок 17. График уменьшения ошибок тенденции точности проверки

В процессе тестирования модели транслятора на каждом этапе проверяется совместимость с данными, содержащимися в параллельном корпусе. Качество перевода улучшается с каждым шагом.

На рисунке 18 представлено тестирование узбекско-казахской NMT модели.

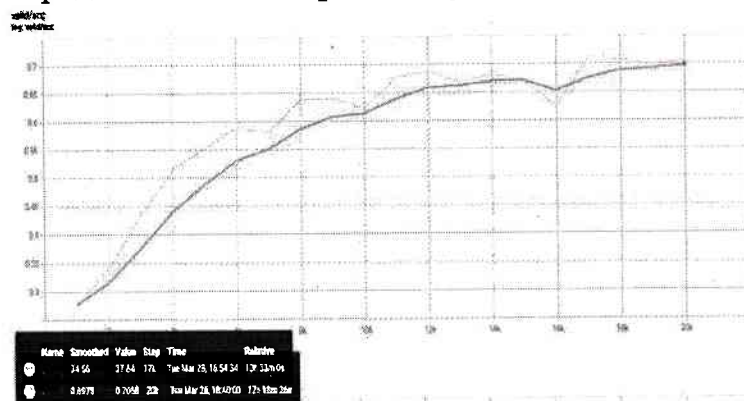


Рисунок 18 - Тестирование узбекско-казахской NMT модели.

Модель перевода была разработана для двух языков и осуществляет перевод с узбекского на казахский и с казахского на узбекский. Поскольку это было сделано впервые, мы не могли сравнить нашу модель с предыдущими работами. Поэтому мы сравнили её с популярными на сегодняшний день системами перевода. В таблице 7 приведены результаты, полученные от узбекско-казахской и казахско-узбекской моделей перевода.

Таблица 7.

Результаты, полученные от модели переводчика:

Предложение на узбекском языке	Яндекс переводчик	Google переводчик	Модель NMT (UzKz)
Unga bera oladigan yordamingiz uchun oldindan rahmat aytmoqchiman.	Мен оған барлық көмегіңіз үшін алдын ала алғыс айтқым келеді.	Оған кез келген көмегіңіз үшін сізге алдын ала алғыс айтқым келеді.	Мен оған көрсете алатын кез келген көмегіңіз үшін сізге алдын ала алғыс айтқым келеді.
O'qituvchi kirib kelganida gapirishni to'xtating.	Мұғалім кірген кезде сөйлесуді тоқтатыңыз.	Мұғалім кірген кезде сөйлесуді тоқтатыңыз.	Мұғалім кірген кезде сөйлесуді тоқтатыңыз.

Предложение на казахском языке.	Яндекс переводчик	Google переводчик	Модель NMT (KzUz)
Ата-анасы сәбиге Акиёши деген ат койды.	Ota-onalar chaqaloq Akiyoshi nomidagi.	Ota-onalar chaqaloqqa Akiyoshi deb nom berishdi.	Ota-onalar chaqaloqqa Akiyoshi deb nom berishdi.
2008 yilda Delta va Northwest aviakompaniyalari birlashib, dunyodagi eng yirik aviakompaniyani tashkil qilishdi.	2008 жылы Delta және Northwest бірігіп, әлемдегі ең ірі авиакомпанияны құрды.	2008 жылы Delta және Northwest Airlines бірігіп, әлемдегі ең ірі әуе компаниясын құрады.	2008 жылы Delta және Northwest әуе компаниялары бірігіп, әлемдегі ең ірі әуе компаниясын құрады.

Из результатов видно, что наша переводческая модель демонстрирует результаты, сравнимые с системами перевода Google и Yandex. Увеличение объема данных позволит достичь еще лучших результатов. Наша переводческая модель это обеспечивает.

ЗАКЛЮЧЕНИЕ

В рамках диссертационной работы на тему «Система машинного перевода между узбекским и казахским языками» получены следующие основные результаты:

1. Изучена грамматика узбекского и казахского языков и определены сходства и различия между ними. Бесконтекстная грамматика Хомского использовалась для формализации синтаксиса простых предложений и было построено дерево компонентов.

2. Сравнительные онтологические модели построены с использованием грамматических правил узбекского и казахского языков. Соответствие между текстами для узбекского и казахского языков согласовывалось на уровне предложений.

3. С использованием параллельного корпуса узбекско-казахского языка сформирована база знаний на основе лингвистических правил. Объем созданной базы знаний составляет 125 тысяч предложений. База знаний размещена на электронной платформе Hugging Face.

4. Разработан нейромашинный перевод для узбекско-казахских языков. Модель «Трансформер» использовалась при разработке машинного перевода. Эта модель позволяет создать систему перевода, которая осуществляет перевод с узбекского на казахский и с казахского на узбекский.

5. Разработана система машинного перевода для узбекского и казахского языков на основе нейронной сети. Система перевода использовалась в качестве помощника при принятии решений экспертами отрасли, позволяя оценить надежность окончательного решения.

6. Точность перевода и семантическая близость определялись с помощью оценки качества текста перевода (BLEU) и показателей взаимной энтропии.

**SCIENTIFIC COUNCIL AWARDING SCIENTIFIC DEGREES
DSc.13/30.12.2021.T.142.01 AT RESEARCH INSTITUTE FOR
DEVELOPMENT OF DIGITAL TECHNOLOGIES AND ARTIFICIAL
INTELLIGENCE**

NATIONAL UNIVERSITY OF UZBEKISTAN

ALLABERDIEV BOBUR BAKHTIYOROVICH

**MACHINE TRANSLATION SYSTEM BETWEEN UZBEK AND KAZAKH
LANGUAGES**

05.01.11 – Digital technologies and artificial intelligence

**ABSTRACT OF DISSERTATION OF THE DOCTOR OF
PHILOSOPHY (PhD) ON TECHNICAL SCIENCES**

Tashkent – 2024

The theme of doctor of philosophy (PhD) on technical sciences was registered at the Supreme Attestation Commission at the Ministry of Higher Education, Science and Innovation of the Republic of Uzbekistan number B2023.4.PhD/T4191.

Dissertation has been prepared at the National University of Uzbekistan named after Mirzo Ulugbek.

Abstract of the dissertation is posted in three languages (Uzbek, Russian, English (resume)) on the website (www.airi.uz) and the "Ziyonet" Information and educational portal (www.ziyonet.uz).

Scientific adviser:

Rakhmonov Zafar Ravshanovich
Doctor of Physical and Mathematical Sciences, docent

Official opponents:

Mukhamedieva Dilnoz Tulkunovna
Doctor of Technical Sciences, Professor

Madrakhimov Shavkat Fayzullaevich
Doctor of Technical Sciences, Professor

Leading organization:

**Tashkent University of Information Technologies,
Nukus Branch**

The defense will take place on "27" december 2024 o'clock at the meeting of Scientific Council No. DSc.13/30.12.2021.T.142.01 at the Research Institute for the Development of Digital Technologies and Artificial Intelligence (Address: 17A, Buz-2, M.Ulugbek district, Tashkent city, 100125/ Phone: (+99871) 263-41-98, e-mail: info@airi.uz).

The dissertation is available at the Information Resource Centre of the Research Institute for Development of Digital Technologies and Artificial Intelligence (is registered under No. 50). (Address: 17A, Buz-2, M.Ulugbek district, Tashkent city, 100125/ Phone: (+99871) 263-41-98).

Abstract of dissertation sent out on "17" december 2024 year.
(mailing report No. 19 on "10" december 2024 year.)



N.Ravshanov
Chairman of Scientific Council
on award of scientific degrees,
Doctor of Technical Sciences, Professor

F.M.Nuraliev
Scientific secretary of Scientific Council
on award of scientific degrees,
Doctor of Technical Sciences, Professor

Sh.KH.Fozilov
Chairman of Scientific Seminar under Scientific
Council on award of scientific degrees,
Doctor of Technical Sciences, Professor

INTRODUCTION (abstract of PhD dissertation)

The aim of the research work is to develop an appropriate model, algorithm, and software package for translating between two Turkic languages using both traditional methods and advanced machine translation models.

The object of research work is a machine translation system for the Uzbek and Kazakh languages.

The scientific novelty of this research includes:

comparative ontological models were built using the grammatical rules of the Uzbek and Kazakh languages;

a knowledge base was formed using ontological models of the Uzbek-Kazakh languages;

taking into account the peculiarities of the Uzbek language, a translation model was created using a neural machine learning algorithm;

a machine translation system for the Uzbek and Kazakh languages was developed based on a neural network.

Implementation of research results. Based on the scientific innovations achieved using the algorithms and software developed for a machine translation system between Uzbek and Kazakh languages, the implementation of research results includes:

identification of similarities and differences between the grammar of the Uzbek and Kazakh languages, creation of a parallel corpus for these two languages, development of ontological models of grammar rules and machine learning algorithms were implemented within the framework of the project of the Samarkand branch of the Tashkent University of Information Technologies named after Muhammad al-Khwarizmi (certificate of the Ministry of Digital Technologies of the Republic of Uzbekistan dated November 20, 2024, No. 33-8 / 7790). As a result, it was possible to increase the volume of the text base of the national corpus of the Uzbek language and improve the accuracy of the software by 15%;

the knowledge base created on the basis of a comparative ontological model of the Uzbek-Kazakh language was successfully implemented in the Samarkand regional branch of Unicon SOFT LLC (certificate of the Ministry of Digital Technologies of the Republic of Uzbekistan dated November 20, 2024, No. 33-8 / 7790). The analysis showed that the translation results of software built on the basis of this knowledge base are comparable to the results of Google and Yandex translation systems. At the same time, it was established that further improvement of the translation quality is possible due to an increase in the volume of data. The introduction of this technology made it possible to reduce the time required for data translation by 11%;

Uzbektelecom was tested and implemented in the process of establishing communication with subscribers of the Samarkand city communication network. As a result of the test, the achievements of the proposed models were compared and the strengths and weaknesses of the system were identified. It was found that the developed models and methods can be used to solve other practical problems of automatic abstraction and processing of natural language texts, as well as in information retrieval systems, curriculum development and e-education systems.

The experiments conducted in research work are explained by the use of any information transfer in education, science, art, healthcare and other social spheres (certificate of the Ministry of Digital Technologies of the Republic of Uzbekistan dated November 20, 2024, No. 33-8 / 7790). As a result, this made it possible to achieve an effective result in the process of communicating with clients at 13.5% speed.

Approbation of research results. Research results were discussed at 6 international and 5 republican scientific and practical conferences.

Structure and volume of the dissertation. The dissertation consists of an introduction, four chapters, a conclusion, a list of references and applications. The volume of the dissertation is 118 pages.

E'LON QILINGAN ISHLAR RO'YXATI
СПИСОК ОПУБЛИКОВАННЫХ РАБОТ
LIST OF PUBLISHED WORKS

I bo'lim (I часть; I part)

1. Allaberdiyev B., Matlatipov G., Kuriyozov E., Rakhmonov Z. Parallel texts dataset for Uzbek-Kazakh machine translation // Data in Brief, 2024. Vol. 5. Art. No.110194. DOI: 10.1016/j.dib.2024.110194 (№3, Scopus IF=0.208). – P. 1-11

2. Allaberdiyev B., Rakhmonov Z., Altynbek Sh., Akhmedova K. Advantages and Disadvantages of Machine Translation Methods // AIP Conference Proceedings, 2023. Vol. 2781. Art. No.020056. DOI: 10.1063/5.0144776 (№3, Scopus IF=0.152). – P. 256-262

3. Allaberdiyev B., Raxmonov Z. Qozoq va o'zbek tillarida gaplarning tuzilishini ontologik modeli // Muhammad Al-Xorazmiy avlodlari Ilmiy-amaliy va axborot-tahliliy jurnal. – T., 2023. – № 3 (25). – B. 12-14 (05.00.00. № 10).

4. Allaberdiyev B. O'zbek-qozoq tillari mashina tarjimasi uchun ma'lumotlar to'plami // Muhammad Al-Xorazmiy avlodlari Ilmiy-amaliy va axborot-tahliliy jurnal. – T., 2024. – № 1 (27). – B. 3-8 (05.00.00. № 10).

5. Allaberdiyev B. O'zbek va qozoq tillarining qisqacha qiyosiy tadqiqi // Yosh tadqiqotchi jurnali. – T., 2022. – № 1 (6). – B. 84-97 (№ 40, ResearchGate IF=4.7).

II bo'lim (II часть; II part)

6. Allaberdiyev B. O'zbek va qozoq tillarning qisqacha qiyosiy morfologiyasi / O'zbekiston Milliy universiteti talabalar va ilmiy tadqiqotchilarining ilmiy konferensiyasi. – Toshkent, 2022. – B. 39-240.

7. Allaberdiyev B. Turkiy tilda sodda gaplarning tuzilishini tadqiq qilish va rasmiylashtirish / II Republican Scientific and Practical Conference of Young Scientists. – Tashkent, 2023. 28-29.03. – P. 178-179.

8. Allaberdiyev B. Form of names of movement through afficium-(i)sh / Of the VII international scientific conference. – Fergana, 2021. 15-17.11. – P. 282.

9. Allaberdiyev B., Raxmonov Z. Brief overview of the grammar of the uzbek language / The international scientific conference of "Contemporary mathematics and its applications". – Tashkent. 2021. 9-21.11. – P. 82.

10. Allaberdiyev B. Qozoq va o'zbek tillari sifatlarining ontologik modeli / The international scientific and practical conference "Actual problems of mathematical modeling and information technology". – Nukus, 2023. 2-3.05. – P. 20-21.

11. Allaberdiyev B. Methods of assembly of parallel corpus for uzbek and kazakh languages / The 8th international conference "Actual problems of applied mathematics and information technologies". – Samarkand, 2023. 25-26.09. – P. 275.

12. Matlatipov S.G., Allaberdiyev B.B., Raxmonov Z.R. O'zbek va Qozoq tillarining parallel korpusi. № BGU 00864. 17.11.2022.

13. Matlatipov S.G., Allaberdiyev B.B. O'zbek va Qozoq tillari uchun mashina tarjima tizimining sun'iy intellect modeli. № DGU 18728. 08.09.2022.

14. Allaberdiyev B.B. Qozoq va o'zbek tillari uchun tarjimon dastur. № DGU 26948. 19.07.2023.

Avtoreferat "O'zMU xabarlari" ilmiy jurnali tahririyatida tahrirdan o'tkazildi.